Multi-channel dereverberation for speech intelligibility improvement in hearing aid applications

Kuklasinski, Adam

DOI (link to publication from Publisher):
10.5278/vbn.phd.engsci.00129

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

MULTI-CHANNEL DEREVERBERATION FOR SPEECH INTELLIGIBILITY IMPROVEMENT IN HEARING AID APPLICATIONS

BY

ADAM KUKLASINSKI

DISSERTATION SUBMITTED 2016

AALBORG UNIVERSITY

DENMARK

Multi-Channel Dereverberation for Speech Intelligibility Improvement in Hearing Aid Applications

Ph.D. Dissertation
Adam Kuklasiński

Dissertation submitted July 15, 2016

Thesis submitted:	July 15, 2016
Ph.D. Supervisor:	Prof. Jesper Jensen Aalborg University and Oticon A/S, Denmark
Assistant Ph.D. Supervisor:	Prof. Søren Holdt Jensen Aalborg University
External Ph.D. Supervisor:	Prof. Simon Doclo Universität Oldenburg, Germany
Ph.D. Committee:	Prof. Emanuël Habets Friedrich-Alexander Universität, Germany Prof. Steven van de Par Universität Oldenburg, Germany Prof. Søren Bech Aalborg University and Bang&Olufsen, Denmark
Ph.D. Series:	Faculty of Engineering and Science, Aalborg University

ISSN: 2246-1248
ISBN: 978-87-7112-762-1

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK-9220 Aalborg Ø
Phone: +45 9940 7140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2016.

About the author

Adam Kuklasiński

Adam Kuklasiński has received the B.Sc. degree in sound engineering and the M.Sc. degree in acoustics from Adam Mickiewicz University, Poznań, Poland, in 2010 and 2012, respectively. He is currently pursuing his Ph.D. degree in digital signal processing at Oticon A/S, Copenhagen, Denmark, and at Aalborg University, Denmark. Mr. Kuklasiński is a Marie Skłodowska-Curie fellow in the Initial Training Network for Dereverberation and Reverberation of Audio, Music, and Speech (ITN-DREAMS). His scientific interests include: statistical signal processing, speech acoustics and enhancement, and binaural sound perception.

Abstract

Excessive reverberation and noise are detrimental for speech quality and can, in severe cases, degrade speech intelligibility. This problem particularly affects individuals with hearing loss. Therefore, it is of practical interest to consider the use of dereverberation and denoising algorithms in digital hearing aids. In this dissertation we propose two such algorithms, one of which is designed for dereverberation and the other one, for joint dereverberation and denoising of speech. The proposed algorithms are based on the well-known multi-channel Wiener filter (MWF) and use maximum likelihood estimators (MLEs) of the power spectral densities (PSDs) of the target speech and the interfering reverberation. In both algorithms, the MWF and the PSD MLEs are derived based on the assumption that the late reverberation is isotropic and is uncorrelated with the target speech, whose direction of arrival (DoA) is known.

By way of numerical simulations the proposed methods are demonstrated to work well in synthetic and in realistically simulated reverberation. Moreover, several instrumental measures indicate that the performance of the proposed algorithms is higher than that of a similar, recently proposed algorithm by Braun and Habets. This result can be explained by the fact that the mean square error of PSD estimation in the proposed algorithms is lower than in the competing algorithm. Nevertheless, comparison of the speech intelligibility resulting from the use of the proposed and the competing algorithm did not reveal statistically significant differences between them.

It is known that the MWF is sensitive to errors in the assumption on the target sound DoA. We investigate this notion by measuring several objective performance scores in a series of simulations where the simulated target DoA differs from the one assumed in the algorithm. The experiments reveal that binaural configurations of the algorithm (i.e. using microphones of the left and the right hearing aid) are far more sensitive to DoA errors than the bilateral configuration (i.e. using microphones of each hearing aid independently). Additionally, we compare the speech intelligibility obtained with the binaural and bilateral implementations of the proposed algorithm under the condition of correct DoA assumption. In this situation, the binaural configuration results in statistically significantly higher intelligibility than the bilateral configuration.

Resumé

Overdreven rumklang og støj skader den opfattede talekvalitet og kan i svære tilfælde føre til nedsat taleforståelse. Dette problem er særligt udtrykt hos personer med høretab. Det er derfor af praktisk interesse at overveje brugen af algoritmer til rumklangs- og støjreduktion i digitale høreapparater. I denne afhandling foreslår vi to sådanne algoritmer, den ene med det formål at reducere rumklang, og den anden med formålet at reducere både rumklang og støj. De beskrevne algoritmer er baseret på det velkendte multikanal-Wienerfilter (MWF) og anvender maximum-likelihood-estimater af de respektive effektspektre for det ønskede talesignal og den forstyrrende rumklang. For begge algoritmer udledes maximum-likelihood estimater og MWF parametre under antagelse af, at rumklangen er isotropisk og uden korrelation til talesignalet, samt at talesignalets ankomstretning er kendt.

Ved brug af numeriske simulationer demonstreres det at de foreslåede metoder fungerer godt i syntetisk og i realistisk simuleret rumklang. Yderligere indikerer adskillige instrumentelle mål, at ydelsen af de foreslåede algoritmer er højere end ydelsen af en lignede metode nyligt foreslået af Braun og Habets. Dette resultat kan forklares ved det faktum, at middelkvadratfejlen ved estimeringen af effektspektre er mindre i de foreslåede algoritmer end for den konkurrerende algoritme. En sammenligning af taleforståelighed ved brug af henholdsvis den foreslåede og den konkurrerende algoritme viste ikke nogen statistisk signifikante forskelle.

Det er velkendt, at et MWF er følsomt over for fejl i den antagne ankomstretning af talesignalet. Vi undersøger dette problem ved at evaluere adskillige objektive performancemål i en række simuleringer, hvor den ankomstretning, der antages af algoritmen, er forskellig fra den faktiske ankomstretning. Disse eksperimenter viser, at binaurale konfigurationer af algoritmen (dvs. konfigurationer, hvor mikrofonsignaler fra både højre og venstre øre anvendes) er langt mere følsomme over for fejl i ankomstretningen end bilaterale konfigurationer (dvs. konfigurationer, hvor der kun anvendes mikrofonsignaler fra et øre ad gangen). Derudover sammenligner vi den opnåede taleforståelighed ved henholdsvist binaurale og bilaterale implementeringer af den foreslåede algoritme under antagelse af, at talesignalets ankomstretning er nøjagtigt kendt. I denne situation giver den binaurale konfiguration anledning til en statistisk signifikant fordel.

Contents

Thesis Details

Thesis Title:	Multi-channel dereverberation for speech intelligibility improvement in hearing aid applications
Ph.D. Student:	Adam Kuklasiński
Supervisors:	Prof. Jesper Jensen, Aalborg University and Oticon A/S Prof. Søren Holdt Jensen, Aalborg University

The main body of this thesis consists of the following five papers:

[A] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," *22nd European Signal Processing Conference (EUSIPCO)*, pp. 61–65, Lisbon, Portugal, 2014.

[B] A. Kuklasiński, S. Doclo, T. Gerkmann, S. H. Jensen, and J. Jensen, "Multi-channel PSD estimators for speech dereverberation – a theoretical and experimental comparison," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 91–95, Brisbane, Australia, 2015.

[C] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1595–1608, 2016.

[D] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Multi-channel Wiener filter for speech dereverberation in hearing aids – sensitivity to DoA errors," *60th Audio Engineering Society International Conference: Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)*, Leuven, Belgium, 2016.

[E] A. Kuklasiński and J. Jensen, "Contralateral microphones in multi-channel Wiener filters for hearing aids – benefits and tradeoffs," submitted to: *Journal of Audio Engineering Society (special issue on Dereverberation and Reverberation of Audio, Music, and Speech – DREAMS).*

Besides the papers included in the main body of the thesis, one additional conference paper has been published:

[81] A. Kuklasiński, S. Doclo, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberant and noisy conditions," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 599–603, Shanghai, China, 2016.

Furthermore, one patent application has been filed:

[67] "Multi-microphone Method for Estimation of Target and Noise Spectral Variances for Speech Degraded by Reverberation and Optionally Additive Noise," J. Jensen and A. Kuklasiński, US20150256956, Sep. 10, 2015.

This thesis has been submitted to the Doctoral School of Engineering and Science at Aalborg University, Denmark, in partial fulfillment of the requirements for a Ph.D. degree. The thesis is based on submitted or published scientific papers which are listed above. Aspects of these papers are used in the thesis introduction. As part of the assessment, co-author statements have been made available to the assessment committee and are also available at the Faculty. The thesis is not in its present form acceptable for open publication but only in limited and closed circulation as copyright may not be ensured.

Preface

This Ph.D. project was carried out in the period between April 2013 and April 2016 within the framework of the Initial Training Network for Dereverberation and Reverberation of Audio, Music, and Speech (ITN-DREAMS). The main part of the work was carried out at Oticon A/S, Copenhagen, Denmark, and some parts of it were done during two secondments: at Aalborg University, Denmark, and at the University of Oldenburg, Germany.

This thesis consists of two parts. Part I serves as an extended introduction and contains background information on speech acoustics and perception (Chapter 1), state of the art review in the field of speech enhancement algorithms (Chapter 2), and a summary of contributions made in this thesis (Chapter 3). Part II constitutes the main body of this thesis and consists of a number of research papers submitted and published throughout the study.

Acknowledgements

The research leading to the results presented in this thesis was financed by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement N° ITN-GA-2012-316969.

* * *

None of what this thesis represents would have been possible without the help, initiative, and professionalism of many excellent persons whom I owe respect and gratitude. The initiator of the ITN-DREAMS project, prof. Toon van Waterschoot, assistant supervisors: prof. Søren Holdt Jensen and prof. Simon Doclo, and, most importantly, my main supervisor, prof. Jesper Jensen, have all contributed to bringing this Ph.D. project to life and to its successful finalization.

I would also like to thank all the wonderful friends I made during the adventure that this Ph.D. project turned out to be. Even more so, I express my gratitude for the never-ending love of my family and for friendships that endured years, as they passed, and transcended distances, as they were covered.

Lastly, I would like to thank my dearest grandparents, late dr. Konrad Dukiewicz and prof. Leokadia Dukiewicz, for transferring onto me their passion of science in general, and speech acoustics research in particular.

Adam Kuklasiński
Copenhagen, July 15, 2016

"Why do you go away? So that you can come back. So that you can see the place you came from with new eyes and extra colors. And the people there see you differently, too. Coming back to where you started is not the same as never leaving."

Terry Pratchett

Part I

Introduction

Introduction

1 Speech communication in noise and reverberation

Speech is, without doubt, one of the most important modes of communication between people. Its importance for establishing oneself in any personal, social, or professional setting can not be overstated. Moreover, exposure to and acquisition of speech is vital for normal development of language, communication and mental skills in children [62, 87]. Thus, speech has deservedly been a subject of scientific inquiry since antiquity. Today, topics such as spoken language structure, its history, speech processing, synthesis, and automatic recognition are each a focal point of separate scientific disciplines [9, 10, 58, 111, 116].

Arguably, speech communication hardly ever takes place in ideal conditions. Indeed, in most cases, the communication channel between the talker and the listener (including the acoustic environment they reside in) is disrupted by noise, reverberation, or competing talkers. At times, speech communication is hindered even further due to the condition of the talker or the listener themselves, e.g. due to speech impediment of the former or hearing impairment of the latter. Motivated by the large number of hearing aid users reporting difficulty communicating in reverberant and noisy environments [77], in this dissertation we undertake the task of mitigating the negative influence of reverberation and noise on speech perception by means of digital signal processing in hearing aids.

In Section 1.1, which directly follows this paragraph, we introduce the basic features of speech signals that are important for speech perception and processing. Next, we describe two categories of frequently occurring interferences: additive noise, in Section 1.2, and room reverberation, in Section 1.3, where we also explain how these interferences affect the perception of speech. We close this chapter with a discussion of some of the mechanisms involved in speech perception by hearing impaired listeners (Section 1.4). In further chapters, we describe the state of the art in speech dereverberation and denoising (Chapter 2) and give an overview of the scientific contributions of this thesis (Chapter 3). Lastly, in Chapter 4, we provide a brief summary of ideas that we consider important topics for future research.

1.1 Signal of interest: speech

As already mentioned, in this thesis the focus is on speech communication scenarios. Therefore, we deem it useful to begin by introducing the most important features of speech signals. For brevity, in this section we describe only the necessary basics. A more complete presentation of the topic of speech production and speech perception can be found e.g. in [40, 61].

In the most general sense, speech is a series of sounds used to encode a message in a language. Speech is generated in a complex process involving coordinated use of one's lungs, vocal cords, tongue, teeth, lips, and other parts of the vocal tract. Interaction of all these elements allows for a wide range of speech sounds, or *phonemes*, to be produced. Similarly to the process of reading that requires the reader to correctly recognize letters, speech understanding requires the listener to correctly recognize phonemes. For example, in English, the words "kill" and "kiss" can be told apart only by recognizing the difference between the phonemes /l/ and /s/ at the end of these words.

Perceptually, individual phonemes are differentiated based on their timbre, pitch, duration, loudness, etc. These perceptual features are linked to various physical characteristics of the sound, which, in turn, are shaped by the configuration of the vocal tract used for production of these phonemes. For example, generation of some phonemes involves vibration of the vocal cords, while for production of other phonemes the vocal cords must remain still. This aspect of articulation determines if the produced sound contains a tonal (i.e. harmonic) component, which enables the listener to differentiate voiced phonemes (i.e. vowels and some consonants, e.g. /b, d, g, z, v/) from voiceless ones (e.g. /p, t, k, s, f/). Many more aspects of articulation must be controlled by the speaker to make a distinction between all the phonemes used in a given language. Correspondingly, in order to recognize these phonemes, the listener must attend to many more qualities of the perceived sounds than just their tonality.

From a signal analysis perspective, speech is a mixture of a wide variety of signal types, some of which are (quasi-)periodic, while other ones are noise-like, or can be described as transients. Thus, the structure of speech signals includes both temporal and spectral features. For this reason, in engineering and for research purposes, speech signals are often represented in terms of their acoustic energy distribution across time and frequency, usually visualized as a spectrogram. In Fig. 1, a spectrogram of an example speech recording is shown along with its waveform and a time-aligned transcription of the recorded phrase. While the waveform in Fig. 1a closely corresponds to the actual physical sound wave received by the microphone, it does not lend itself to easy visual analysis. The spectrogram in Fig. 1c is far more useful for this purpose because it reveals the underlying structure of different phonemes. With training, phonemes and words can be recognized without ever listening to the speech recording—a technique once proposed to allow the deaf to understand speech [106].

Many important features of speech signals can be observed in Fig. 1. First and foremost, *speech is a time-varying signal*: the envelope of the waveform in

Fig. 1a varies dynamically across time, as does the power spectral density in the spectrogram in Fig. 1c. Nevertheless, across short spans of time (∼40 ms), the general speech features reflected in the spectrogram appear to be approximately constant. This can be observed in Fig. 2, where an enlarged portion of the plots from Fig. 1 is shown. Another important feature of speech signals that is visible in Fig. 1 is the fact that they are *a mixture of quasi-periodic and noise-like spectro-temporal regions*. All vowels and many consonants are voiced, and, therefore, contain a harmonic component. The vibration of the vocal cords present in the sound of these voiced phonemes is visible as regular pulses in the waveform in Fig. 2a. The corresponding harmonic structure is visible as parallel horizontal lines in the spectrograms. Voiceless phonemes lack the harmonic structure and can be recognized by their noise-like appearance in the spectrograms. The final observation from Fig. 1 that we wish to make is that *repetitions of any given sequence of phonemes generally result in different signal realizations*. This can be observed by comparing the two similar, but nevertheless different, occurrences of the sequence "china" in Fig. 1. One may conclude that phonemes can not be described in terms of specific values of signal parameters, but rather in terms of (presumably contiguous) regions in the space spanned by these parameters. Individual realizations of a given phoneme are subject to random variation in the production process, but are also influenced by the neighborhood of other phonemes in the utterance (i.e. co-articulation), intonation and stress due to the sentence structure, and even by the emotional state of the speaker [40]. Naturally, gender, age, and individual differences between speakers also have a considerable impact on speech characteristics.

Different languages, and even different dialects of the same language, use different sets of phonemes. For example, some languages use many vowels (Danish [6]) while others rely more heavily on consonants (Polish [33]). Thus, in many experimental paradigms, results obtained using one language are typically not generalizable to other languages. For this reason, when recorded speech is used as a research tool, it is beneficial to use speech material that is diverse in terms of language, or at least contains speech of different talkers. For example, the TIMIT database [41] contains speech of many native, male and female users of the American English language and, thus, partially fulfills this requirement. Another recording that is useful for research purposes is the international speech test signal (ISTS) [59]—a one-minute-long unintelligible speech-like mixture composed of short segments of speech in different languages. One can argue that spectral and temporal characteristics of the ISTS are to some extent representative of speech signals in general, which makes it suitable for use in exploratory and comparative experiments. Despite the aforementioned considerations, in some cases, speech material read by a single person in only one language is used, e.g. in speech intelligibility tests such as Dantale [122].

We close this section by noting that speech is highly redundant and can be understood even after much of the original signal has been removed. For example, old analogue telephone systems were sufficient for communication

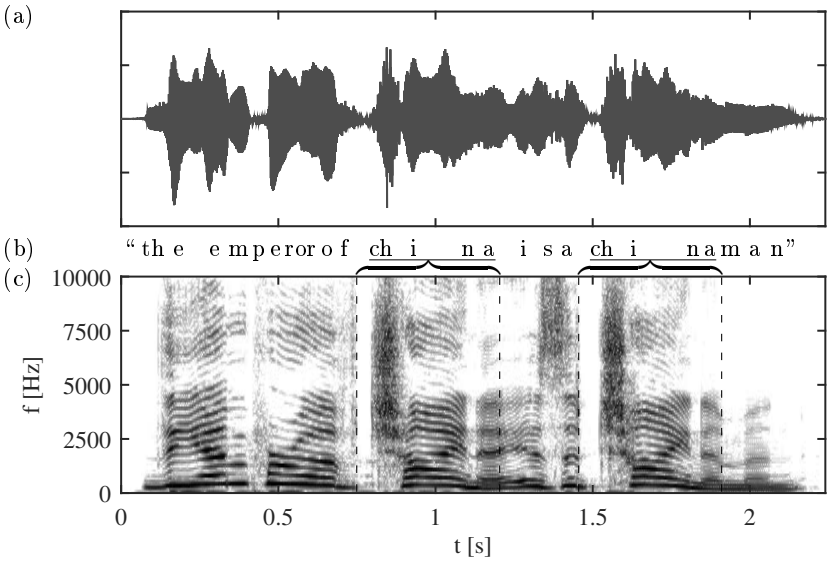


Fig. 1: An example speech recording of the phrase: “The Emperor of China is a Chinaman” in the form of its: *(a)* waveform, *(b)* transcription, and *(c)* spectrogram. The text is time-aligned with the waveform and the spectrogram. The two occurrences of the sequence “china” are underlined, allowing for their comparison in terms of signal characteristics.

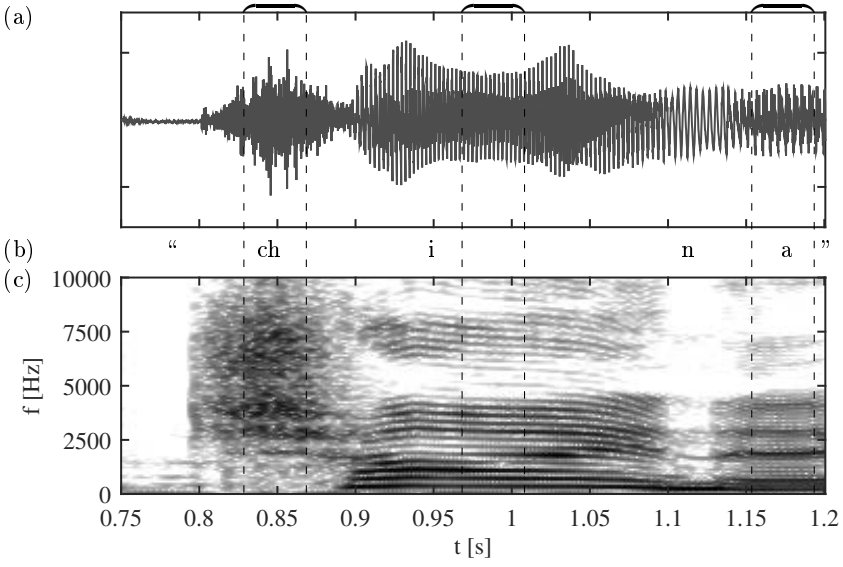


Fig. 2: Enlarged section of the plots from Fig. 1 corresponding to the first occurence of the sequence “china”. The voiceless sound “ch” has a noise-like appearance. The harmonic structure of the voiced phonemes “i-n-a” is visible as (quasi-)periodic pulses in the waveform and as a harmonic structure in the spectrogram. Three fragments where the signal may be considered approximately stationary are indicated. (Each of them is 40 ms long.)

purposes although they transmitted only a part of the full frequency spectrum of speech [63]. As we describe in detail in Sections 1.2 and 1.3, the speech signal is also highly robust to noise and reverberation.

1.2 Interference: noise and competing talkers

Speech communication often takes place in noisy conditions. Working, traveling, socializing, and many other important activities in a person's life are inevitably accompanied by noise. Additionally, in many situations, competing talkers are active at the same time as the desired talker. All these additional sounds that are present in an acoustic scene interfere with the target speech, decreasing its perceived quality and, in more severe cases, its intelligibility [103]. Moreover, interferences increase the listening effort that the listener must invest in the perception process. This may result in fatigue, causing discomfort and loss of focus, decreasing productivity during meetings and classes [44, 56].

The impact of noise on speech perception can be explained by a mixture of at least three distinct effects: energetic masking, informational masking, and binaural advantage. We describe them in the following paragraphs.

Energetic masking arises when the sound of interest is rendered inaudible because of another, co-occurring sound (referred to as the "masker"). Energetic masking is frequency-specific, i.e. it is greatest when the sound of interest and the masker are of the same frequency and diminishes when these two sounds are distant in frequency. This behavior is due to the characteristics of the human hearing and, more specifically, of the basilar membrane in the cochlea [95, 104]. Masking can also occur when the target sound and the masker directly follow one after the other. This is referred to as temporal masking [95].

Evidently, energetic masking has significant implications for noisy speech perception. As an example, consider an acoustic scenario where stationary broadband noise (e.g. as in a car cabin) interferes with the target speech. Depending on the level of the noise, some parts of the speech signal may be masked. The quiet phonemes (e.g. consonants such as /p, f, n, m, h/) are masked more easily than the loud ones (i.e. vowels). This can be demonstrated using the spectrogram in Fig. 3 where a signal composed of the speech recording from Fig. 1 and a stationary pink noise interference is presented. Comparing the spectrograms in Fig. 1 and Fig. 3, it is easy to notice that in Fig. 3 many of the quieter phonemes from the original signal are buried in noise and that their spectral features are no longer discernible to the eye. It is reasonable to expect that some of these phonemes would also be masked. Obviously, the higher the level of the noise, the greater proportion of the phonemes is masked.

Although listeners are often able to recover the message even from partially masked speech, this becomes increasingly difficult for decreasing signal-to-noise ratios (SNRs). The value of the broadband SNR for which 50% of spoken words are identified correctly is an often used measure of speech recognition performance referred to as the speech reception threshold, or the SRT. Besides the level, spectral and temporal characteristics of the masker may also have

an influence on speech perception [108]. For example, in scenarios where the masker's level is time-varying, the listeners are sometimes able to reconstruct the message from the glimpses of speech they perceive during the low-noise-level segments of the utterance [26]. This ability, however, depends also on the listener's knowledge of the language and on the context [70].

Informational masking is related to the inability of the listener to focus their attention on the speech of the desired talker in the presence of a distracting or disorienting sound, especially a competing talker. The effects of energetic and informational masking usually occur together, in which case the term "informational masking" is used to refer to the loss of intelligibility in excess of what can be attributed to the energetic masking [19, 27]. Unlike in energetic masking, the extent of informational masking depends not only on the physical properties of the masker, but also on its "content", or "meaning". For example, time-reversed speech or speech in a language that is not understood by the listener has been shown to be a less effective masker than speech in a known language presented normally [27, 107].

Acoustic conditions with multiple competing talkers are frequently encountered during everyday activities and, consequently, are of particular practical interest. There exists a large body of research work (likely started by Cherry in 1953 [20]) focusing on speech perception in multi-talker interference (for an overview see [19]). From this research it is understood that the "cocktail party problem" (as it is often referred to) involves not only energetic and informational masking effects, but also other mechanisms, e.g. binaural processing, which we describe next.

Binaural advantage is a perceptual benefit arising from the exploitation of the fact that the sounds received by the left and the right ear are generally different. In acoustic scenarios where the target and the interference originate from different locations around the listener, the binaural advantage enables the listeners to obtain detection thresholds or SRTs that are much lower than with the use of just one ear. This ability relies on a number of cues, the most important of which are the interaural level differences (ILDs) and the interaural time differences (ITDs) [12]. By exploiting these cues, for specific positions of the target sound source and a single point-source interferer, listeners are able to detects sounds of levels as low as 15 dB [12] lower than without the binaural advantage. An analogous phenomenon has been observed in speech intelligibility tests. In [11], for a frontally positioned target speaker and a single point-source interferer positioned at 100° off-center, listeners were able to achieve SRTs 12 dB lower than when both sources were placed frontally (i.e. without the benefit of binaural processing).

1.3 Interference: room reverberation

Besides noise and competing talkers, reverberation is likely the third most often encountered type of acoustic interference. Reverberation occurs whenever a sound is emitted in an enclosed space and, thus, is extremely common because

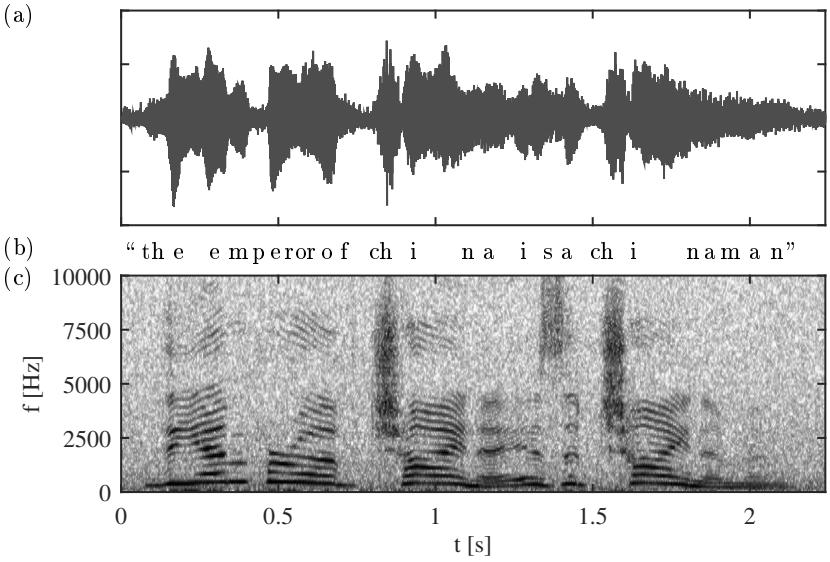


Fig. 3: *(a)* Waveform and *(c)* spectrogram of the speech signal from Fig. 1 with pink noise (A-weighted SNR = 13 dB). Many of the quieter phonemes are below the noise floor and can not be discerned in the spectrogram (an example of energetic masking).

of the amount of time many people spend inside buildings.

Reverberation arises due to repeated reflections of a sound from walls and other surfaces in a room. The sound, after it is emitted by the source, may travel many times across the room before its energy dissipates due to absorption in the reflecting surfaces and in the air. Because of the finite speed of sound, the decay of the acoustic energy sometimes takes a noticeable amount of time (but rarely more than a few seconds). It is exactly this persistence of the acoustic energy in a room that is the essence of reverberation. The rate at which the acoustic energy decays in a given room is an important parameter and is usually expressed as the *reverberation time*, or T_{60}, which is defined as the time that it takes for the acoustic energy to decay a million times (i.e. 60 dB) [82]. Another important parameter of reverberation is the direct-to-reverberation ratio (DRR), which is defined as the ratio of the acoustic energy reaching the receiver directly from the source to the total energy of the reflected sound. Unlike T_{60}, which is approximately constant across different locations in a room, the DRR strongly depends on the distance between the source and the receiver [82].

Reverberation is the superposition of all reflections arriving at a given location in the room. In other words, reverberation is composed of many delayed and attenuated copies of the original sound. It follows, that at any given moment, a reverberant space may be considered to be a linear system and, thus, its influence on the sound signal may be modeled by a convolution with an

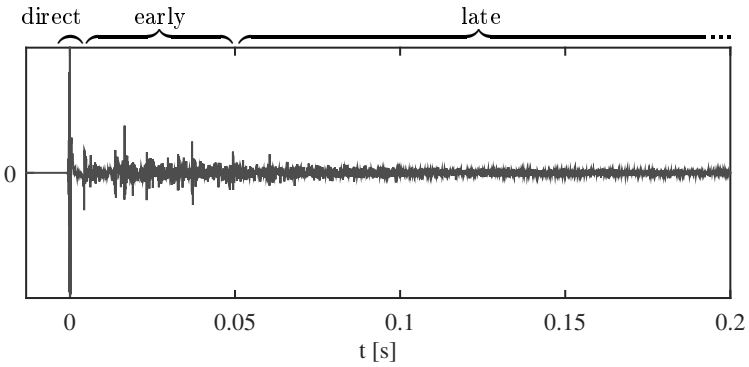


Fig. 4: A typical room impulse response. Direct path response, early reflections, and late reverberation are labeled accordingly.

impulse response. Room impulse responses (RIRs) are useful representations of the acoustic properties of rooms and are often measured for technical and research purposes. It is important to note that RIRs correspond to specific source and receiver positions within a room. Thus, even a small deviation in the source/receiver position may result in a very different RIR [96]. Moreover, fluctuations in the air temperature also have an impact on the RIRs [36]. For these reasons, when a RIR is repeatedly measured, each time a slightly different response is usually obtained.

An example of a measured RIR is presented in Fig. 4. Three important parts of that RIR are indicated: direct path response, early reflections, and late reverberation. The direct path response is always the first impulse in a RIR and is usually the strongest. This is because the line-of-sight is the shortest possible path from the source to the receiver and, thus, results in the least attenuation and delay. The part of RIRs that follows the direct path response is referred to as the early reflections. In typical RIRs, such as in Fig. 4, early reflections are visible as fairly distinct spikes that continue up to 50–100 ms after the onset of the RIR [82]. The last part of RIRs is referred to as the late reverberation and is characterized by a noise-like appearance and approximately exponential decay (barely visible on the time and amplitude scale of Fig. 4) [82].

The three parts of RIRs have specific characteristics with respect to perception of reverberant speech. The direct path speech is considered to be perceptually the most valuable part of the reverberant speech signal because it is essentially a delayed and attenuated version of the original speech signal emitted by the talker [15]. By definition, the direct path speech arrives at the listener's location from the direction that the speaker is physically located at. This is obviously important for speaker localization by the listener.

The early reflections are usually considered advantageous for speech perception because listeners integrate them into the same auditory percept as the direct speech [15]. Nevertheless, in some rooms, early reflections may cause

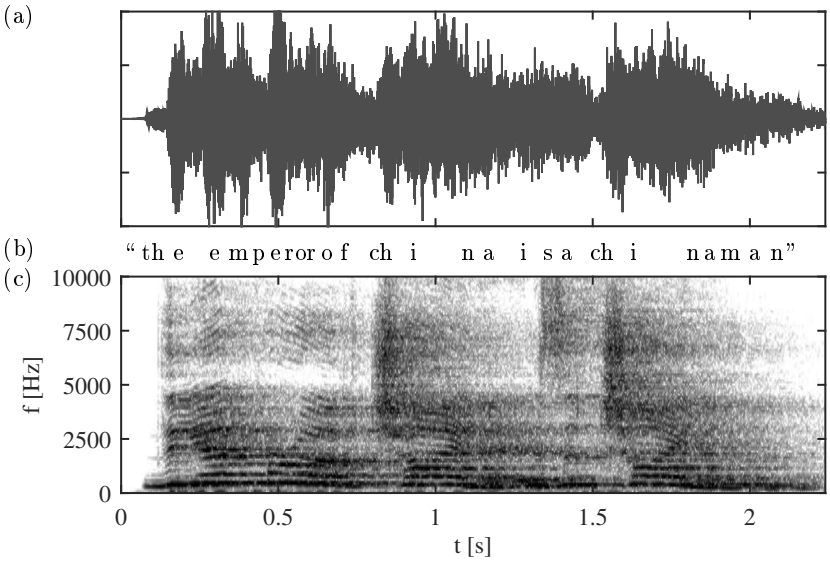


Fig. 5: *(a)* Waveform and *(c)* spectrogram of the speech signal from Fig. 1 in reverberation ($T_{60} = 1\,s$, $DRR = 0\,dB$). Many of the quiet phonemes are masked by the reverberation tail evoked by the preceding louder phonemes (an example of overlap-masking).

perceivable coloration of the speech signal [82]. Although early reflections are likely to arrive from different directions than the direct speech, they normally do not interfere with the speaker localization. This phenomenon is referred to as the precedence effect [123] (the sound that arrives first usually determines the perceived location of its source).

The part of the reverberation that is detrimental for speech intelligibility is the late reverberation. This can be explained e.g. by the fact that the speech energy arriving later than 50 ms after the direct sound is likely to overlap with the next phoneme in a speech utterance. For some phoneme combinations, such as a loud vowel followed by a quiet consonant, this can result in energetic masking of the quiet phoneme by the decay of the energy related to the preceding, louder phoneme (so-called overlap-masking [14]). This effect is clearly visible in Fig. 5, where a synthetically reverberated version of the speech utterance from Fig. 1 is shown. Unlike the direct path and the early reflections, the late reverberation does not have a clearly defined direction of arrival. Instead, the late reverberant sound field is distributed across all possible directions [45]. In the absence of more detailed information, this spatial energy distribution is often assumed to be uniform, or isotropic. This corresponds to the assumption of reverberation diffuseness made in statistical room acoustics (see e.g. [82]).

The fact that the direct and early speech (i.e. the useful signal component) and the late reverberation (i.e. the interference) arrive at the listener's position from different directions creates a potential for obtaining a binaural

advantage. Indeed, binaural listening has been found to significantly improve speech intelligibility in reverberation, especially when noise is also present (see e.g. [11, 18, 92]). Interestingly, the presence of reverberation decreases the binaural advantage in acoustic scenarios with a point-like target source and competing speech sources, compared to analogous but anechoic conditions [92].

1.4 Speech perception by hearing impaired subjects

As we explained in the last two sections, excessive noise and reverberation can make speech understanding a difficult task for any listener. In this section, we focus on speech perception by listeners with hearing impairment, who are affected by adverse acoustic conditions to even greater extent than normal hearing subjects. Because the majority of cases of hearing impairment is due to so-called cochlear hearing loss [93], we consider only this type of hearing impairment.

The main cause of cochlear hearing loss is the damage to inner and outer hair cells (IHCs/OHCs) that reside in the organ of Corti inside the cochlea. This damage may be caused by aging, excessive noise exposure, specific drugs and toxins, or it may be congenital [34, 94]. Because the IHCs and the OHCs are vital for the cochlea to function properly, their loss has far-reaching consequences for the listener [93]: inability to perceive quiet sounds (loss of sensitivity), decreased dynamic range and abnormal loudness perception, inability to discriminate between co-occurring sounds (loss of spectro-temporal selectivity), decreased ability to benefit from binaural listening, and more. Loss of sensitivity and abnormal loudness perception can, in principle, be remedied by using hearing aids with signal amplification and dynamic range compression. In contrast to that, loss of spectro-temporal selectivity and decrease in binaural processing advantage are, even conceptually, more difficult to mitigate.

Because in this thesis we are concerned with speech communication scenarios in noise and reverberation, spectro-temporal selectivity of the hearing and binaural listening advantage are of particular relevance. Thus, the work presented in this thesis is aimed towards mitigating the influence of the loss of these hearing abilities. Because reduced hearing sensitivity and phenomena related to loudness perception can be considered less important for reverberant and noisy speech perception, we assume that they have already been compensated for, e.g. by a dynamic range compression system as usually found in hearing aids [52, 93].

In noisy and reverberant speech communication scenarios, reduced spectro-temporal selectivity and binaural processing ability of the impaired hearing generally results in: *(a)* higher SRT in noise, *(b)* decreased ability to segregate the target speech from competing voices, and *(c)* reduced robustness against reverberation. This may lead to increased listening effort, difficulties in communication, and, as a potential consequence, in social isolation. Therefore, in speech enhancement systems for hearing aids the goal is frequently to: *(a)* suppress the background noise, *(b)* suppress the sound of competing speakers

while preserving the target speech component, and *(c)* reduce the amount of reverberation. Ideally, this should facilitate normal communication and decrease the listening effort.

2 Enhancement of reverberant and noisy speech

In hearing aids, but also in other speech communication applications such as hands-free telephony and voice controlled devices, the target speaker is almost always at a considerable distance from the microphone(s) of the receiving/recording device (in the range of meters). Because of this distance, the speech-to-interference ratio of the received signal is typically much lower than in close-microphone applications such as traditional telephony (where the distance between the target source and the microphone(s) is in the range of centimeters). This is caused by the fact that ambient noise and (late) reverberation have approximately the same level across all positions in a room, whereas the target speech level decreases as the distance from the target speaker increases.

To enhance the noisy and reverberant speech signal, noise reduction and/or dereverberation algorithms are used. These have the goal of restoring speech quality, intelligibility, or automatic speech recognition performance. In hearing aids, speech intelligibility improvement is of greatest interest. Moreover, certain speech enhancement algorithms have been shown to reduce the listening effort [109], which is also beneficial for the hearing aid users. In the literature, many types of processing algorithms have been proposed for speech dereverberation and/or noise reduction in many applications, including hearing aids. In this chapter, we provide an overview of the state of the art in this area.

Most of the existing speech denoising and dereverberation algorithms fall into one or more of the following categories [86, 98]: *(a)* spectral processing algorithms, *(b)* spatial processing algorithms, or *(c)* system identification and inversion algorithms. For clarity, we discuss these categories in separate sections, focusing on algorithm types applicable to hearing aids.

2.1 Spectral processing

Spectral-based speech enhancement algorithms operate on a spectro-temporal representation of the noisy/reverberant speech, typically, by manipulating the short-time Fourier transform (STFT) coefficients of the input signal. Spectral-based algorithms are used for noise and reverberation reduction based on the assumption that the target speech and the interference are differently distributed in the spectro-temporal plane. This allows for selective attenuation of the spectro-temporal regions dominated by the interference. Processing paradigms range from purely heuristic, through statistical model-based, and up to recent methods based on e.g. neural networks or non-negative matrix factorization.

Heuristic methods of speech spectral enhancement include several classes of algorithms based on different assumptions and spectrum modification tech-

niques (for an overview see e.g. [85]). For example, the long-established method of spectral subtraction is based on the idea that enhanced speech may be obtained by subtracting an estimate of the magnitude spectrum of the interference (or a power thereof) from that of the input signal. In the literature, there exist many spectral subtraction algorithms for speech enhancement in noise (e.g. [13, 71]) and in reverberation (e.g. [48, 83]). While heuristic speech spectral enhancers provide some level of interference reduction, they are generally considered inferior to algorithms based on statistical models and optimality criteria [55], whose description follows.

Statistical model-based spectral enhancement algorithms use an *a priori* statistical model of the input signal STFT and its components. Algorithms are derived by minimizing a chosen cost function given the assumed statistical model of the signal (for overviews, see [8, 86]). In statistical model-based spectral enhancers, input signal STFT coefficients are usually assumed to be statistically independent across frequency and time. Moreover, the target and the interference are most often assumed to be mutually statistically independent, which allows the power spectral density (PSD) of the input signal to be modeled as the sum of the PSDs of individual signal components. Most algorithms based on the aforementioned assumptions differ in: *(a)* the type of the cost function used in the derivation, *(b)* the assumed probability distribution of the speech component of the STFT, *(c)* the assumed probability distribution of the interference component of the STFT, and *(d)* the manner in which the parameters of the speech and interference statistical models are estimated.

One of the most often used cost functions is the minimum mean square error (MMSE) of the complex-valued STFT. If a circularly-symmetric complex Gaussian distribution of target and interference STFT coefficients is assumed, optimization with respect to this cost function results in the well-known Wiener filter. Instead of the MMSE of the complex-valued STFT coefficients, other cost functions are also sometimes used. For example, optimization of the MMSE of the short-time spectral amplitude (STSA) under target and interference Gaussianity assumption leads to the well-known STSA-MMSE algorithm by Ephraim and Malah [37]. It is important to note that the assumption that the speech STFT coefficients are complex Gaussian distributed is only approximately correct. In fact, is has been demonstrated that speech STFT coefficients are much better described by super-Gaussian distributions, particularly by the Laplace distribution [42, 66] or the Gamma distribution [39, 89]. Using these super-Gaussian distributions to model the speech (and sometimes noise) STFT coefficients leads to algorithms that are somewhat more complicated than the Wiener filter or the Ephraim-Malach algorithm, but can result in an improved speech enhancement performance [39, 89].

Generally speaking, spectral enhancement algorithms depend on the knowledge of statistical parameters of the input signal, of which at least some are not known and have to be estimated from the noisy/reverberant observations. For example, in many noise reduction spectral enhancement algorithms, the noise PSD is one of these required parameters. Based on the assumption

that the noise is approximately stationary, minima-tracking noise PSD estimators are frequently used, for example the well-known minimum statistics method [88], or later methods such as [21, 43, 54]. Besides the noise PSD, the signal SNR (i.e. the ratio of the target speech and the interference PSDs) is also required by most spectral enhancement algorithms[1]. In many cases, this quantity is referred to as the *a priori* SNR. A classical method for *a priori* SNR estimation is the decision-directed estimator by Ephraim and Malah [37]. Newer and more advanced methods are also used, e.g. [17, 22, 24]. Importantly, in speech dereverberation algorithms, the interference can not be assumed to be stationary or slowly-evolving compared to the target speech PSD (as it is in many spectral-based noise reduction algorithms). Thus, in spectral-based speech dereverberation algorithms, specialized late reverberation PSD estimators are used, many of which rely on the assumption that the late reverberant energy decays exponentially [105], e.g. [50, 51, 83].

Relatively recently, spectral enhancement algorithms departing from the usually employed assumption of statistical independence of the STFT coefficients across time and frequency have been proposed. These algorithms operate on patches of the STFT, which allows them to exploit spectro-temporal dependencies—an obvious advantage for processing of structured signals such as speech (see Section 1.1). Some methods are based on the non-negative matrix factorization (NMF) technique and model target speech and interference spectrograms as two different low rank matrices whose sum equals the input signal spectrogram [72, 91]. Methods based on supervised learning, e.g. deep neural networks (DNNs), are also increasingly common, e.g. [53, 124].

Besides the aforementioned speech enhancement algorithms operating in the STFT domain, other algorithms using other signal transforms (e.g. the Karhunen–Loève transform, often implemented using the singular value decomposition) also exist. These algorithms have the aim of enabling separation of the signal into the target and interference subspaces and subsequent projection of the noisy observation onto the target subspace [38, 68, 84].

Spectral enhancement algorithms are primarily intended for use in single microphone systems, but some of them can also be used with multiple microphones. For example, the spectral-based reverberation reduction algorithms in [49, 51] can be used with one, but also with many microphones, which improves these algorithms' performance. Similarly to many other multi-microphone speech enhancement algorithms, these algorithms are composed of a spatial pre-processor and a spectral enhancement scheme. We describe more of this class of enhancement algorithms together with spatial processing algorithms in the following section.

[1] It should be noted that, given a noise PSD estimate, *a priori* SNR estimation is effectively equivalent to target speech PSD estimation.

2.2 Spatial processing

Spatial-based speech enhancement algorithms work by jointly processing and combining signals of an array of microphones. This allows for realization of systems that have a spatial selectivity pattern that is different (and usually more directional) than that of any of the individual microphones. Spatial-based algorithms are used for speech enhancement based on the assumption that the target speech and the interferences impinge on the microphone array from different spatial regions. This allows for selective attenuation of the sounds impinging from directions other that of the target speech.

One of the simplest methods of spatial processing is to time-shift the microphone array signals such that sounds impinging on the array from a chosen target direction are aligned. Subsequently, these microphone signals are added together, causing coherent summation of the target sound (speech) and incoherent summation of sounds arriving from other directions. Fittingly, this method is referred to as the delay-and-sum beamformer (DSB) [118]. Although the DSB optimally reduces interference that is uncorrelated between microphones (e.g. microphone self-noise), it is not optimal for reduction of point noise sources or diffuse interference such as late reverberation. This problem can be solved by using more advanced, signal-dependent beamformers. For example, minimum variance distortion-less response (MVDR) and linearly constrained minimum variance (LCMV) beamformers [28] use target and interference statistics to optimally reduce the interference while preserving the target signal with a pre-specified gain. Many more beamforming techniques exist, a comprehensive overview of which can be found e.g. in [119].

As previously mentioned, beamformers are frequently used in combination with spectral enhancement schemes, which results in a two-step algorithm. In this context, the spectral enhancement step is referred to as a post-filter because it is typically applied after the spatial processing step. Among the first methods of this type were heuristic algorithms for speech dereverberation and noise reduction such as [4] and [127], both composed of a DSB and a coherence-based post-filter. Similarly to the DSB, these post-filters work best in scenarios where the target speech component is coherent between the microphones and the interference is spatially white. However, in realistic acoustic scenarios the interference is often generated by a point-source or is diffuse and, therefore, exhibits some degree of spatial correlation, particularly at low frequencies [35]. This limits the speech enhancement performance of coherence-based post-filters and the DSB. Nevertheless, coherence-based post-filters have an important advantage over the single-channel, minimum-statistics-based spectral enhancement algorithms described earlier: they are capable of adapting to the interference level variations not only during speech absence, but also during speech activity. Thanks to this feature, coherence-based post-filters are suitable for reverberation reduction in speech signals without making any assumptions on the reverberant energy decay.

Limitations of the DSB and the coherence-based post-filters can be over-

come by using algorithms that are based on optimality criteria and a statistical model of the signal. Arguably, one of the most frequently used statistically-motivated spatial enhancement algorithms for speech processing is the multi-channel Wiener filter (MWF) [30–32]. Similarly to the single channel Wiener filter described in the previous section, the MWF is derived by optimizing the MMSE cost function under the assumption that the signal components are Gaussian. In acoustic scenarios where the target speech is generated by a single point-source, the MWF can be factored into an MVDR beamformer and a single channel Wiener post-filter [112]. This decomposition of the MWF is frequently used in practical applications because it allows for the beamformer and the post-filter to be controlled and monitored separately.

As mentioned earlier, the MVDR beamformer depends on certain statistics of the input signal components. Specifically, required are a so-called target steering vector and the inter-microphone covariance matrix of the interference. In acoustic scenarios where the spatial features of the target and the interference are time-invariant, these required statistics can also be assumed to be time-invariant and, sometimes, even to be known *a priori*. In other applications, on-line estimators of the target steering vector (or, sometimes equivalently, the DoA) [23, 115] and the covariance matrix of the interference [110] must be employed. The second part of the MWF—the spectral Wiener post-filter—depends on the *a priori* SNR or, equivalently, on the time-varying target and interference PSDs. Estimation of these PSDs from multiple microphone signals in reverberant and noisy conditions using the maximum likelihood methodology is one of the main focus areas in this thesis.

2.3 System identification and inversion

Unlike the speech dereverberation algorithms based on spectral or spatial features of the signal, system identification and inversion algorithms exploit the convolutive nature of the reverberation. More specifically, algorithms of this type attempt to estimate the RIR (system identification) and apply a filter that equalizes it (system inversion). Compared to the already described classes of speech dereverberation algorithms, system identification and inversion algorithms are unique in that they, theoretically, can achieve perfect dereverberation [90], provided that multiple microphone signals are available (and some additional conditions).

The research in the area of speech dereverberation by system identification and inversion has started several decades ago [90, 99]. Of particular importance is the contribution in [90], where the multiple-input/output theorem (MINT) is postulated. Many challenges in practical application of the system identification and inversion algorithms have been encountered, but significant progress has been made since. Specifically, challenges related to RIR invertibility and sensitivity to erroneous RIR estimates and additive interference have been important focus points, e.g. in [57, 97].

As explained in Section 1.3, from the point of view of speech intelligibility,

reduction of the late reverberation is of particular interest. This fact has been one of the motivations behind an important class of system identification and inversion algorithms that instead of equalizing the RIR, attempt to shorten it (so-called channel shortening) or to reduce the energy of the late reverberation while preserving the direct and early sound energy, e.g. [128]. More recently, important contributions in this area were made, in particular the partial MINT (PMINT) by Kodrasi [78–80].

2.4 Special considerations related to hearing aids

Speech enhancement algorithms are increasingly often used in state-of-the-art hearing aids [52]. However, several important aspects of this specific application must be considered for the hearing aid and for the enhancement algorithm to serve their respective purposes without interfering with each other. We describe some of these aspects in the following paragraphs.

Functional blocks of a typical hearing aid

All hearing aids contain at least one input transducer (i.e. a microphone), an output transducer (i.e. an earphone[2]), and a signal processing circuit. In the vast majority of state-of-the-art hearing aids the processing circuit is a digital signal processor (DSP). In it, usually at least three main functionalities are implemented [52, 93]: hearing loss compensation, feedback reduction, and signal enhancement. A diagram of the signal flow between processing blocks corresponding to these functionalities is depicted in Fig. 6.

Arguably, the most important functionality of hearing aids is the *hearing loss compensation*, which is usually realized as a dynamic range processor with a compressive characteristic [93]. This allows quiet sounds to be amplified more than the loud ones, such that the hearing aid user can perceive a wide dynamic range of sounds at a comfortable, yet audible, level.

Depending on the severity of the hearing loss, hearing aids might apply a significant, time- and frequency-dependent amplification of the input signal. This, in connection with small distances between the microphones and the earphone, is likely to result in acoustic feedback between the output and the input of the hearing aid. In result, whistling or howling sounds are produced, especially when a hand or a telephone is held against the ear. To avoid this undesirable phenomenon, *feedback reduction* algorithms such as frequency shifting, adaptive notch filtering, or adaptive feedback cancellation are often used in hearing aids [52]. A systematic overview of the state of the art in the area of feedback cancellation may be found in [120] (for general applications) and in [46] (for application in hearing aids).

As described in Section 1.4, hearing impairment not only results in inaudibility of quiet sounds, but also in increased difficulty in speech understanding

[2]Hearing care professionals usually refer to the hearing aid's output transducer as the "receiver".

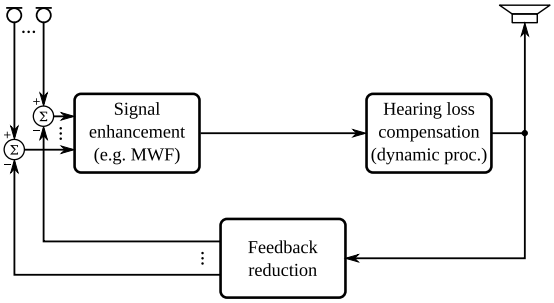


Fig. 6: Simplified diagram of the signal flow in a state-of-the-art hearing aid.

in noisy conditions. To mitigate this problem, *signal enhancement* algorithms are often used in hearing aids. For the reasons explained in Section 1.4, in this thesis we neglect the hearing loss compensation and the feedback reduction processing blocks, and focus on speech enhancement algorithms for hearing aid applications.

Technical requirements for speech enhancement algorithms used in hearing aids

Several technical requirements exist which impose limitations on the speech enhancement algorithms that are used in hearing aids. For example, the signal delay introduced by the hearing aid must be kept to a minimum. Otherwise, several types of problems might arise. If the delay is larger than approximately 80 ms, problems with respect to synchronousness of visual and auditory stimuli can arise [113]. This is particularly detrimental for the hearing impaired as they rely on lip-reading more than normal hearing listeners [25]. However, even shorter delays can interfere with speech perception by the hearing aid user. This is so because many hearing aids use so-called open fittings, which allow the unprocessed, airborne sound to enter the ear canal. Thus, the sound reaching the ear drum and perceived by the user is a combination of the unprocessed and the processed (and delayed) sound. This creates a potentially disturbing comb filtering effect which must be avoided. For these reasons, signal processing algorithms used in hearing aids cannot introduce delays that exceed a few, normally not more than ten, milliseconds.

The second technical limitation imposed by the application in hearing aids is the amount of processing power available for use in these small and battery powered devices. While technical advancements in semiconductor manufacturing continue to make this limitation more and more relaxed, computational complexity of processing algorithms is an important factor that must be considered in hearing aid design.

Lastly, unlike many other electronic devices (e.g. smartphones) that are de-

signed to actively engage their users, hearing aids must operate without much, or preferably any, user interaction. Thus, speech enhancement algorithms used in hearing aids must, ideally, operate correctly under all possible acoustic conditions or, alternatively, automatic detection of relevant acoustic scenarios must be implemented and used to select different processing schemes according to the changing acoustic conditions.

Distortion of spatial cues

As mentioned in Section 1.2, human's ability to localize and separate sound sources relies on a number of cues. Clearly, it is of interest to ensure that hearing aids do not distort these cues, or at least that the introduced distortions do not interfere with the spatial awareness of the hearing aid user.

Some of the most important spatial cues are the already mentioned binaural cues [12]: ILDs and ITDs. Hearing aids, due to the fact that they change the level and introduce phase differences in the processed signals, generally distort these cues. Perfect preservation of binaural cues is nearly impossible to achieve in hearing aids, but certain steps may be taken to minimize ITD and ILD distortions. For example, hearing loss compensation normally involves signal amplification with a time-varying, signal-dependent gain which, in general, is different in each hearing aid[3]. Clearly, this results in ILD distortions. However, if the same time- and frequency-dependent gain could be applied to the signals in both hearing aids, ILD distortions could be minimized. In fact, in some hearing aids it is possible to use a wireless communication channel between the left and the right device to exchange the necessary information and synchronize the applied gains. Apart from amplifying the signal, hearing aids also introduce phase differences in the processed signals. The resulting ITD distortions can be reduced if the two hearing aids introduce the same amount of latency and perform the same operations on the signal (e.g. the same type of beamforming).

Certain spatial cues (particularly the ones relevant for sound localization in the vertical plane) are not related to interaural differences, but rather to spectral features that stem from the shape of the pinna [12]. Unfortunately, these cues may be lost when behind-the-ear hearing aids are used. Although this problem theoretically could be solved by using in-the-ear or in-the-canal hearing aids, i.e. hearing aids whose microphones are located at the entrance of the ear canal, the experimental results in [101] suggest otherwise. This is in contrast to a more recent study where state-of-the-art hearing aids of behind-the-ear, in-the-ear, and in-the-canal types have all been found to allow for near-normal sound localization in the vertical plane [29]. Clearly, a full understanding of spatial perception in the case of aided, impaired hearing is yet to be reached by the research community.

[3] In this thesis we assume that the user is fitted with two hearing aids, one on each ear.

Binaural and bilateral hearing aids

In most cases, hearing aids work independently of each other or with only little information, such as program or volume settings, synchronized wirelessly between them (so-called bilateral configuration). Future hearing aids are expected to make use of a "binaural link" – a digital transmission channel allowing the hearing aids to exchange information about the microphone signals in real time (resulting in a so-called binaural configuration) [52]. As described above, this can be used for minimizing ILD distortions. Moreover, the binaural link can be used to transmit microphone signals between the hearing aids, enabling their joint processing by a signal enhancement algorithm. This is known to increase the performance of certain spatial processing algorithms such as the MWF [117].

2.5 Evaluation of speech enhancement algorithms

It is not a trivial task to compare and rate the performance of speech enhancement algorithms in a meaningful way. This is because there is no single definition of what constitutes a "good" speech enhancement algorithm. In more formal words, there exist many potential performance criteria that can be used to evaluate and compare speech enhancement algorithms. In this section, we give a brief overview of the most often used performance measures applicable to speech enhancement algorithms.

In this thesis, the focus is on speech enhancement algorithms intended for improvement of speech intelligibility (or quality) as attained (or perceived) by human listeners. Speech enhancement algorithms are also used for other purposes, e.g. as a front-end in automatic speech recognition systems. However, these algorithms and the performance measures used for their evaluation are outside of the scope of this overview.

Arguably, *listening tests with human subjects* are the most direct and reliable method of evaluating speech enhancement algorithms whose output is intended for presentation to human listeners. Listening tests are a versatile research tool because they enable a wide range of important subjective and objective parameters to be measured. Subjective parameters, such as speech quality or the perceived amount of reverberation, are usually measured using the mean opinion score (MOS) or the multiple stimuli with hidden reference and anchor (MUSHRA) methodologies. Objective parameters, such as speech intelligibility, are measured using other methods, e.g. the diagnostic rhyme test [1], digit triplets tests [64, 102], or sentence tests with [121, 122] or without [100] a fixed grammatic structure. Besides subjective and objective evaluation of perceptual qualities of sounds, experimental paradigms for indirect measurement of the listening effort also exist, e.g. [109, 126]. Despite their versatility and reliability, listening tests with human subjects have several disadvantages: they are time-consuming, require the availability of sufficient (potentially hearing impaired) subjects, and their results may not always be easy to interpret.

In order to circumvent the inconveniences related to listening tests with human subjects, many *predictors of listening test results* have been proposed. These predictors are normally based on a computational model of a given aspect of auditory perception, e.g. perceived speech quality or intelligibility. The perceptual evaluation of speech quality (PESQ) algorithm [2] and its successor, the perceptual objective listening quality assessment (POLQA) [3, 7], are some of the most often used speech quality predictors. Speech intelligibility predictors also exist, e.g. the short time objective intelligibility measure (STOI) [114] and its later extensions [65] that improve its reliability in acoustic scenarios with modulated interference [69] or enable it to predict the influence of binaural effects [5]. Specialized speech intelligibility and quality predictors for hearing aids, HASPI and HASQI, respectively, have been proposed by Kates in [73, 74].

The above-mentioned instrumental performance measures are convenient to use because they can be relatively quickly calculated, e.g. based on a series of computer simulations of the acoustic scenario and the enhancement algorithms that are of interest. However, instrumental performance measures should be used with care as the range of acoustic situations and processing types for which they can produce reliable results is not necessarily well-known and is certainly limited. Although computational predictors of listening test results are not as versatile or reliable as real listening tests, the fact that they can be relatively easily computed for a large number of test conditions makes them particularly suitable for exploratory research. Ultimately, a real listening test is usually conducted to validate and confirm the predictions of the model.

Besides perceptual measures and estimators thereof, *technical performance measures* may also be useful and may provide important insights into diferences between algorithms. Unlike the predictors of listening test results treated above, the technical performance measures aim at characterizing technical aspects of the evaluated algorithms. It should be noted, however, that the difference between predictors of listening test results and technical performance measures is not clean-cut and some measures share characteristics of both classes. For example, SNR improvement is a simple, useful, and often used technical performance measure. However, to account for the fact that different frequencies are of different importance for speech perception, frequency-weighting of the SNR improvement is often used [60]. Technical performance measures typically compare the output of the signal processing algorithm or communication device in question with an undistorted, unprocessed reference signal. Hence, these measures are mostly useful in laboratory situations where a clean reference signal is readily available. Other technical performance measures allow for separate evaluation of the interference reduction performance and the amount of distortions introduced into the speech component of the signal (see e.g. the measures defined in [39, 47]). For more examples of technical performance measures see [60, 75].

3 Summary of contributions

The main topic of the work presented in this thesis is speech dereverberation in hearing aids. Because in many acoustic scenarios speech is distorted not only by reverberation but also by noise, we focused on algorithms that can be used for joint reduction of these two types of interferences. As described in Chapter 2, many classes of speech dereverberation and denoising algorithms exist in the literature. Due to the requirements imparted by the application in hearing aids, we considered only the algorithms that can operate with low input/output latency. Moreover, out of consideration for battery life and due to physical limitations in the computing power available for use in hearing aids, we focused on algorithms whose complexity allows for implementation on modern or near-future hearing aid platforms.

Amongst the algorithm classes mentioned in Chapter 2, spectral- and spatial-based algorithms fit well with the requirements of the hearing aid application. They both can be implemented as on-line or block-based processing schemes. Moreover, they tend to be of lower computational complexity than system identification/inversion algorithms. To achieve maximum performance, in this thesis we focused on two-step algorithms where a spatial pre-processor is combined with a spectral post-filter, allowing us to exploit both the spectral and spatial structure of reverberant and noisy speech signals.

A general diagram depicting the structure of the type of algorithms considered in this thesis is shown in Fig. 7. Notably, in this structure, the spectral post-filter is based on the *multi-channel input* of the spatial pre-processor (indicated by the the dotted lines in Fig. 7) as opposed to the *single channel output* of this pre-processor. Thus, the class of algorithms that we considered is more than just a concatenation of a spatial- and a spectral-based processing algorithm. Moreover, we considered *binaural* hearing aids, i.e. composed of an interconnected pair of hearing devices, one on each of the user's ears. This enabled us to evaluate the expected benefit of binaural hearing aids over traditional bilateral configurations (without the binaural link).

The choice of the specific type of the spatial pre-processor and the spectral post-filter was driven by our assumptions on speech, reverberation, and noise Gaussianity and their mutual statistical independence. In such a signal scenario, minimization of the mean square error leads to the multi-channel Wiener filter (MWF) which, under some additional conditions, can be factored into an MVDR spatial beamformer and a single channel spectral Wiener post-filter. For implementation of the MWF, knowledge of the target speech and the interference (noise and reverberation) inter-microphone covariance matrices is required. These are not normally known, but they can be estimated if an appropriate statistical model of the reverberant and noisy speech signal is employed.

In the work presented in this thesis, we assume that reverberation is isotropic (i.e. it is evenly distributed across all direction around the hearing aid user) and that the target speech is at a known and time-invariant location with re-

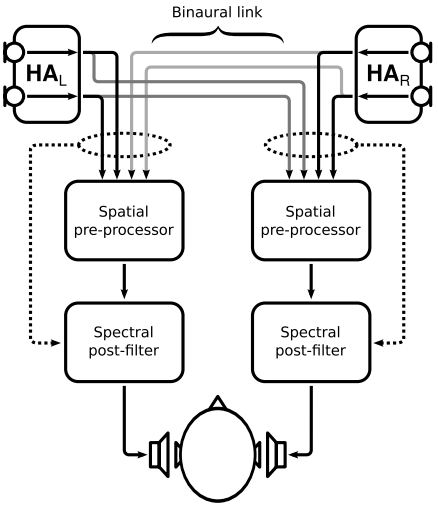


Fig. 7: Block diagram of the class of speech enhancement systems considered in this thesis. In binaural hearing aids, the microphone signals are exchanged between the left and the right device. In a bilateral hearing aid, the binaural link is absent and the two devices work independently of each other.

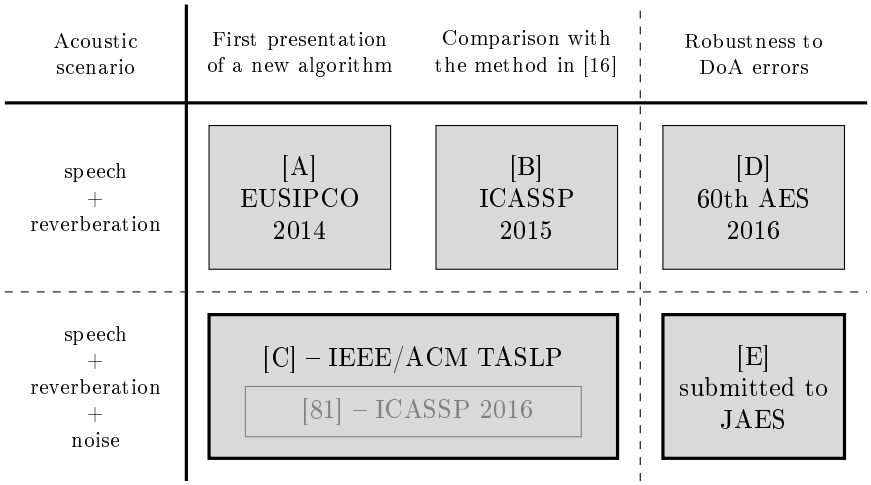
spect to the user. This allows us to formulate a statistical signal model suitable for derivation of the MWF. As we describe in more detail in the main body of this thesis (particularly in Paper A and C), estimation of the target speech and the reverberation power spectral densities (PSDs) is crucial for implementation of the MWF. Much of the contributions of this thesis pertain to the problem of estimation of these parameters.

3.1 Scope of contributions

The main body of this thesis (i.e. Part II) consists of a collection of research papers. In this section we summarize contributions made in each of them and outline how they are related to each other.

The first three papers are primarily devoted to the problem of estimation of the target speech and the late reverberation PSDs in the signal scenario described above. Overall, we proposed two estimators of the target speech and the late reverberation PSDs. The first estimator is applicable to a simplified signal scenario without the additive noise component and is presented in Paper A and further evaluated in Paper B. This simplified model of the reverberant speech signal allowed the use of a pre-existing maximum likelihood estimator of the speech and reverberation PSDs. The second PSD estimator is proposed in Paper C and is a generalization of the estimator from Paper A for acoustic conditions that include additive noise. This resulted in a more realistic, but

Table 1: Table summarizing relations between the collected papers. Rows indicate different acoustic scenarios and columns indicate different scopes of the experiments considered in individual papers. Journal papers are outlined with a thicker line.



also more complicated, model of the input signal. In effect, a novel PSD MLE needed to be derived for this purpose.

The last two of the collected papers explore the performance of the MWFs based on the two proposed estimators in an acoustic scenario which was found to be particularly problematic for the proposed algorithms: when the assumed target speech direction of arrival (DoA) does not correspond exactly to the actual DoA. Sensitivity to this error is highly dependent on the use of the binaural/bilateral hearing aid configuration. For this reason, we have included four different microphone array configurations in our experiments. In Paper D and in Paper E we evaluate the robustness to DoA mismatch of the algorithms from Paper A and Paper C, respectively. Moreover, in Paper E we compared a binaural and a bilateral configuration of the MWF from Paper C in terms of speech intelligibility.

For clarity, we summarize the relations between the collected papers in Table 1. *The acoustic scenarios* considered in individual papers are determined by the rows of the Table 1 and *the types of experiments* are indicated by its columns. In the following paragraphs we provide a detailed description of contributions made in the individual papers.

[A] "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids"

In Paper A, we propose an MWF-based speech dereverberation algorithm which uses a maximum likelihood estimator of the signal components' PSDs. The

MWF and the PSD estimator are based on the assumption that the microphone signal has only two components: cylindrically isotropic reverberation and target speech that is generated by a point-source. This assumption results in a particularly simple MWF structure where the beamforming part is constant. Additionally, for this simple signal model, the MLE of the target speech and the reverberation PSDs is mathematically identical to estimators from [76, 125], which are of closed form.

Through a series of simulations, we demonstrate that the proposed algorithm is able reduce reverberation in both synthetic and realistic reverberant conditions. Moreover, we show that, as could be expected, a 4-microphone, binaural hearing aid configuration of the algorithm outperforms a 2-microphone, single hearing aid version.

[B] "Multi-channel PSD estimators for speech dereverberation – a theoretical and experimental comparison"

In Paper B, we compare the algorithm proposed in Paper A with another MWF-based speech dereverberation algorithm proposed by Braun and Habets in [16]. The algorithm in [16] is based on a similar set of assumptions as employed in Paper A. However, unlike in Paper A, in [16] an additive noise component is included in the signal model. To enable a meaningful comparison of the algorithms, in this paper we assume that this noise component is absent.

Besides the signal model, the only significant difference between the compared algorithms is in the used speech and reverberation PSD estimators. Thus, initially, our comparison focuses on these PSD estimators. Through numerical simulations as well as analytical derivations, we show that for microphone arrays of more than two microphones the PSD estimator from Paper A outperforms the estimator from [16] in terms of estimation accuracy. Additionally, analytical derivations allow us to conclude that for arrays of two microphones the PSD estimators form Paper A and [16] are identical.

For completeness, we compare the speech dereverberation performance of the MWFs based on the two PSD estimators. The results indicate that the better estimation accuracy of the method in Paper A leads to a small advantage in terms of dereverberation performance of the MWF, as measured by objective performance measures, compared to the algorithm in [16].

[C] "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise"

In Paper C, we propose a novel algorithm that is a generalization of the algorithm from Paper A for acoustic scenarios that, besides target speech and reverberation, also include additive noise. This results in a more realistic signal model, but it necessitates derivation of a novel PSD estimator. The resulting estimator is not of closed form and results in a higher overall computational complexity of the proposed algorithm compared to the algorithms in Paper A

and [16].

Through two numerical experiments we show that for arrays of more than two microphones: *(a)* the proposed PSD estimator achieves higher estimation accuracy than the estimator used in [16], and *(b)* that the speech dereverberation performance of the MWF using the proposed estimator is somewhat higher than that obtained using the algorithm form [16]. As in the noise-free scenario considered in Paper B, the two algorithms are identical when only two microphones are in use. In addition to numerical simulations, we conduct a speech intelligibility test with 20 subjects. The results indicate similar speech intelligibility improvements over the unprocessed signal when using either of the algorithms.

Paper C builds upon a conference paper [81] entitled "Maximum likelihood PSD estimation for speech enhancement in reverberant and noisy conditions" which was presented at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) in 2016. This preliminary investigation lacks the PSD estimation accuracy comparison and the objective evaluation of the algorithms' performance includes fewer performance measures than in Paper C. Moreover, the speech intelligibility results presented in [81] are based on only the first 10 subjects of the listening test.

[D] "Multi-channel Wiener filter for speech dereverberation in hearing aids – sensitivity to DoA errors"

Paper D is a preliminary study of the robustness of the algorithm proposed in Paper A to erroneous target DoA assumption. Four different microphone array configurations are compared in terms of their absolute performance and robustness to DoA errors. Binaural configurations of the algorithms are shown to be capable of higher dereverberation performance but also that they are much more sensitive to the DoA error. We explain this through an analysis of directional sensitivity patterns of the beamforming part of the MWF in the four microphone array configurations.

[E] "Contralateral microphones in multi-channel Wiener filters for hearing aids – benefits and tradeoffs"

Paper E is an extension of the study in Paper D for acoustic scenarios that include additive noise (besides target speech and reverberation). As in Paper D, we test and compare four microphone array configurations in terms of their absolute performance and robustness to DoA errors. The experiment in this paper includes a more comprehensive set of true and assumed DoAs than in Paper D, which provides additional insights into the relative performance of different microphone configurations. Besides the numerical simulations with instrumental performance measures, we conduct a speech intelligibility test with 20 subjects. In the test included are two microphone array configurations of the MWF from Paper C: a bilateral one (two hearing aids working independently) and a bin-

aural one (two interconnected hearing aids operating on microphone signals from both sides). The results indicate that under correct DoA assumption, the binaural configuration of the MWF results in a statistically significantly better speech intelligibility than the bilateral configuration.

To conclude, binaural configurations of the algorithm in Paper C can result in higher dereverberation performance at the cost of much higher sensitivity to DoA errors.

3.2 Summary of conclusions

The overall conclusions of the work presented in this thesis can be divided into three themes related to: *(a)* the MLEs of target speech and late reverberation PSDs, *(b)* the use of the MLE-based MWF for speech dereverberation in hearing aids, and *(c)* the advantages and disadvantages of using specific microphone array configurations with MWFs in hearing aids.

The proposed MLEs of target speech and late reverberation PSDs generally perform well in the acoustic scenarios considered in this thesis. More specifically, in noise-free, as well as in noisy reverberant single-talker scenarios, the proposed MLEs performed better than the competing estimator by Braun and Habets [16]. The proposed estimators also appear to be more robust to violation of the reverberation isotropy assumption than [16]. However, in the noisy reverberant speech scenario the estimator proposed in Paper C is more computationally complex than the estimator from [16].

The use of the proposed MLE-based MWFs for speech dereverberation in hearing aids appears to be beneficial, as shown by their high performance in terms of instrumental measures, as well as speech intelligibility measured in a listening test with human subjects. Compared to the method in [16], the proposed algorithm achieves higher FWSegSNR and PESQ scores. However, the two methods are difficult to distinguish perceptually and result in similar speech intelligibility improvement.

Binaural configurations of the MWF, while providing increased speech dereverberation performance in idealized conditions, are shown to be very sensitive to errors in the assumed target DoA. Mismatch as small as 15° between the assumed and the true DoA resulted in a steep decline in the performance scores of the binaural MWF, whereas the bilateral configuration performed equally well for a wider range of DoA errors.

4 Directions for future research

While the proposed MLE-based MWF algorithm generally performed well in the experimental conditions that were tested in this work, several areas for further research have been identified. Moreover, new questions arose in the aftermath of this work.

The first area that we deem worthwhile for future research is the continued work on extending the signal model used in the proposed algorithms. Explicit modeling of the target speech early reflections, uncertainty associated with the target speech source direction of arrival, competing talkers—all these extensions would widen the applicability of the algorithm. Extension with an on-line estimator of the target speech steering vectors appears particularly important in light of the high sensitivity to steering mismatch, discovered in Papers D and E.

Due to the low-latency requirement necessary in hearing aid applications, all results reported in the collected papers were obtained with a relatively short STFT frame length of 8 ms. Consequently, the frequency resolution of the proposed algorithms was lower than it could have been had a more conventional frame length of 32 ms been used instead. Moreover, the use of short STFT frames excluded the possibility of incorporating early reflections into target steering vectors. Clearly, it is if interest to investigate the performance of the proposed algorithms for applications where a larger latency can be tolerated, such that STFT frame lengths in the order of tens of milliseconds could be used.

The use of spatial processing algorithms in hearing aids, particularly in binaural hearing aids, affects the binaural cues that listeners use for correct localization and separation of sound sources. Clearly, extending the proposed algorithms with binaural cue preservation is of high relevance.

In a longer time-frame, we may expect a continued growth of the processing power of integrated circuits that are used in hearing aids. This may enable practical use of more advanced signal processing paradigms such as deep neural networks (DNNs) or other methodologies which rely on more accurate models of the acoustic scenario, the human listener, or both.

References

[1] "Method for measuring the intelligibility of speech over communication systems," *ANS S3.2-1989*, 1989.

[2] "Perceptual evaluation of speech quality: an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Rec. P. 862*, 2001.

[3] "Perceptual objective listening quality assessment (POLQA)," *ITU-T Rec. P. 863*, 2011.

[4] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912–915, 1977.

[5] A. H. Andersen *et al.*, "Predicting the intelligibility of noisy and non-linearly processed binaural speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2016, (early access, DOI: 10.1109/TASLP.2016.2588002).

[6] H. Basbøll, *The Phonology of Danish*, ser. The Phonology of the World's Languages. OUP Oxford, 2005.

[7] J. G. Beerends *et al.*, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement, parts I and II," *J. Audio Eng. Soc*, vol. 61, no. 6, pp. 366–384 and 385–402, 2013.

[8] J. Benesty, S. Makino, and J. Chen, "Introduction," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Berlin, Germany: Springer, 2005, ch. 1, pp. 1–8.

[9] J. Benesty, M. M. Sondhi, and Y. Huang (Eds.), *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

[10] M. Benzeghiba *et al.*, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007.

[11] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 331–342, 2006.

[12] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[13] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr 1979.

[14] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Acoust. Soc. Am.*, vol. 21, no. 6, pp. 577–580, 1949.

[15] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am.*, vol. 113, no. 6, pp. 3233–3244, 2003.

[16] S. Braun and E. A. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. 21st Eur. Signal Process. Conf. (EUSIPCO)*, Marrakech, Morocco, 2013, pp. 1–5.

[17] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, March 2008, pp. 4897–4900.

[18] C. Breitsprecher, "Effects of reverberation on speech intelligibility in normal-hearing and hearing-impaired listeners," Master's thesis, Technical University of Denmark, 2011.

[19] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[20] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, 1953.

[21] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sept 2003.

[22] ——, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 870–881, Sept 2005.

[23] ——, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, 2004.

[24] ——, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *Signal Process.*, vol. 86, no. 4, pp. 698–709, Apr. 2006.

[25] R. Conrad, "Lip-reading by deaf and hearing children," *British Journal of Educational Psychology*, vol. 47, no. 1, pp. 60–65, 1977.

[26] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[27] M. Cooke, M. L. Garcia Lecumberri, and J. Barker, "The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception," *J. Acoust. Soc. Am.*, vol. 123, no. 1, pp. 414–427, 2008.

[28] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 35, no. 10, pp. 1365–1376, 1987.

[29] T. V. den Bogaert, E. Carette, and J. Wouters, "Sound source localization using hearing aids with microphones placed behind-the-ear, in-the-canal, and in-the-pinna," *Int. J. Audiology*, vol. 50, no. 3, pp. 164–176, 2011.

[30] S. Doclo, "Multi-microphone noise reduction and dereverberation techniques for speech applications," Ph.D. dissertation, Katholieke Universiteit Leuven, 2003.

[31] S. Doclo *et al.*, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7-8, pp. 636–656, Jul.–Aug. 2007.

[32] ——, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. R. Liu, Eds. Wiley, 2008, pp. 269–302.

[33] L. Dukiewicz and I. Sawicka, *Fonetyka i fonologia.* Instytut Języka Polskiego PAN, 1995, vol. 3, [in Polish].

[34] C. Elberling and K. Worsoe, *Fading sounds: about hearing and hearing aids.* The Oticon Foundation, 2005. [Online]. Available: http://www.fadingsounds.com

[35] G. W. Elko, "Spatial coherence functions for differential microphones in isotropic noise fields," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001, pp. 61–85.

[36] G. W. Elko, E. Diethorn, and T. Gänsler, "Room impulse response variation due to temperature fluctuations and its impact on acoustic echo cancellation," in *Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Kyoto, Japan, 2003, pp. 67–70.

[37] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[38] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul 1995.

[39] J. S. Erkelens *et al.*, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.

[40] J. L. Flanagan, *Speech analysis synthesis and perception*. Springer Science & Business Media, 2013, vol. 3.

[41] J. S. Garofolo *et al.*, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*. NIST, 1993.

[42] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 204–207, JUL 2003.

[43] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[44] P. A. Gosselin and J.-P. Gagné, "Older adults expend more listening effort than young adults recognizing speech in noise," *J. Speech, Language, Hearing Res.*, vol. 54, no. 3, pp. 944–958, 2011.

[45] B. N. Gover, J. G. Ryan, and M. R. Stinson, "Measurements of directional properties of reverberant sound fields in rooms using a spherical microphone array," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2138–2148, 2004.

[46] M. Guo, "Analysis, design, and evaluation of acoustic feedback cancellation systems for hearing aids – a novel approach to unbiased feedback estimation," Ph.D. dissertation, Aalborg Universitet, Denmark, 2013.

[47] S. Gustafsson *et al.*, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 245–256, Jul 2002.

[48] E. A. P. Habets, "Single-channel speech dereverberation based on spectral subtraction," in *15th Annu. Workshop Circuits, Systems, Signal Process.*, 2004, pp. 250–254.

[49] ——, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 4, March 2005, pp. 173–176.

[50] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–773, Sept 2009.

[51] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, 2007.

[52] V. Hamacher *et al.*, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 2915–2929, Jan. 2005.

[53] K. Han, Y. Wang, and D. Wang, "Learning spectral mapping for speech dereverberation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2014, pp. 4628–4632.

[54] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, March 2010, pp. 4266–4269.

[55] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art*, ser. Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool Publishers, 2013, vol. 9, no. 1.

[56] C. B. Hicks and A. M. Tharpe, "Listening effort and fatigue in school-age children with and without hearing loss," *J. Speech, Language, Hearing Res.*, vol. 45, no. 3, pp. 573–584, 2002.

[57] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, pp. 1–12, 2007.

[58] W. Holmes, *Speech synthesis and recognition*. CRC press, 2001.

[59] I. Holube *et al.*, "Development and analysis of an international speech test signal (ISTS)," *Int. J. Audiology*, vol. 49, no. 12, pp. 891–903, 2010.

[60] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan 2008.

[61] X. Huang, A. Acero, and H.-W. Hon, "Spoken language structure," in *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001, pp. 19–72.

[62] J. R. Hurford, "The evolution of the critical period for language acquisition," *Cognition*, vol. 40, no. 3, pp. 159–201, 1991.

[63] A. H. Inglis, "Transmission features of the new telephone sets," *Transactions of the American Institute of Electrical Engineers*, vol. 57, no. 10, pp. 606–612, 1938.

[64] S. Jansen *et al.*, "The french digit triplet test: A hearing screening tool for speech intelligibility in noise," *Int. J. of Audiology*, vol. 49, no. 5, pp. 378–387, 2010.

[65] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2016, to appear.

[66] J. Jensen, I. Batina, R. C. Hendriks, and R. Heusdens, "A study of the distribution of time-domain speech samples and discrete fourier coefficients," in *Proc. SPS-DARTS*, vol. 1, 2005, pp. 155–158.

[67] J. Jensen and A. Kuklasiński, "Multi-microphone method for estimation of target and noise spectral variances for speech degraded by reverberation and optionally additive noise," Sep. 10, 2015, US Patent 20150256956.

[68] S. H. Jensen *et al.*, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 6, pp. 439–448, Nov 1995.

[69] S. Jørgensen, R. Decorsière, and T. Dau, "Effects of manipulating the signal-to-noise envelope power ratio on speech intelligibility," *J. Acoust. Soc. Am.*, vol. 137, no. 3, pp. 1401–1410, 2015.

[70] D. N. Kalikow, K. N. Stevens, and L. L. Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *J. Acoust. Soc. Am.*, vol. 61, no. 5, pp. 1337–1351, 1977.

[71] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 4, Orlando, FL, USA, 2002, p. 4164.

[72] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, April 2009, pp. 45–48.

[73] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[74] ——, "The hearing-aid speech quality index (HASQI) version 2," *J. Audio Eng. Soc*, vol. 62, no. 3, pp. 99–117, 2014.

[75] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The reverb challenge: Acommon evaluation framework for dereverberation and recognition of reverberant speech," in *IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, Oct 2013, pp. 1–4.

[76] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Bucharest, Romania, 2012, pp. 295–299.

[77] S. Kochkin, "Marketrak V: 'why my hearing aids are in the drawer': The consumers' perspective." *The Hearing Journal*, vol. 53, no. 2, pp. 34–36, 2000.

[78] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 680–693, April 2016.

[79] I. Kodrasi, "Dereverberation and noise reduction techniques based on acoustic multi-channel equalization," Ph.D. dissertation, Carl von Ossietzky Universität Oldenburg, 2016.

[80] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1879–1890, 2013.

[81] A. Kuklasiński, S. Doclo, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberant and noisy conditions," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, 2016, pp. 599–603.

[82] H. Kuttruff, *Room Acoustics*, 5th ed. Taylor & Francis, 2009.

[83] K. Lebart, J. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[84] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Process. Lett.*, vol. 10, no. 4, pp. 104–106, April 2003.

[85] P. C. Loizou, "Spectral-subtractive algorithms," in *Speech Enhancement: Theory and Practice*. Taylor & Francis, 2007, ch. 5, pp. 97–139.

[86] ——, *Speech Enhancement: Theory and Practice*. Taylor & Francis, 2007.

[87] A. R. Luria and F. I. Yudovich, *Speech and the Development of Mental Processes in the Child*, J. Simon, Ed. Penguin Books, 1971.

[88] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[89] ——, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, 2005.

[90] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, no. 2, pp. 145–152, 1988.

[91] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2140–2151, Oct 2013.

[92] J. P. Moncur and D. Dirks, "Binaural and monaural speech intelligibility in reverberation," *J. Speech Hearing Res.*, vol. 10, no. 2, pp. 186–195, 1967.

[93] B. C. J. Moore, "Speech processing for the hearing-impaired: successes, failures, and implications for speech mechanisms," *Speech Communication*, vol. 41, no. 1, pp. 81–91, 2003.

[94] ——, "Physiological aspects of cochlear hearing loss," in *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues.* John Wiley & Sons, Ltd, 2008, pp. 1–37.

[95] ——, *An introduction to the psychology of hearing.* Brill, 2012.

[96] J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *J. Sound and Vibration*, vol. 102, no. 2, pp. 217–228, 1985.

[97] J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 7. Paris, France: IEEE, 1982, pp. 1858–1861.

[98] P. A. Naylor, "Introduction," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. Springer, 2010.

[99] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, no. 1, pp. 165–169, 1979.

[100] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1085–1099, 1994.

[101] W. Noble and D. Byrne, "A comparison of different binaural hearing aid systems for sound localization in the horizontal and vertical planes," *British J. Audiology*, vol. 24, no. 5, pp. 335–346, 1990.

[102] E. Ozimek *et al.*, "Development and evaluation of Polish digit triplet test for auditory screening," *Speech Communication*, vol. 51, no. 4, pp. 307–316, 2009.

[103] K. L. Payton, R. M. Uchanski, and L. D. Braida, "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," *J. Acoust. Soc. Am.*, vol. 95, no. 3, pp. 1581–1592, 1994.

[104] J. O. Pickles, *An introduction to the physiology of hearing.* Brill, 2012.

[105] J.-D. Polack, "Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics," *Applied Acoustics*, vol. 38, no. 2-4, pp. 235–244, 1993.

[106] R. Potter, G. Kopp, and H. Green, *Visible speech*, ser. Bell Telephone Laboratories series. D. Van Nostrand Co., 1947.

[107] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Release from informational masking by time reversal of native and non-native interfering speech," *J. Acoust. Soc. Am.*, vol. 118, no. 3, pp. 1274–1277, 2005.

[108] ——, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, 2006.

[109] A. Sarampalis *et al.*, "Objective measures of listening effort: Effects of background noise and noise reduction," *J. Speech, Language, Hearing Res.*, vol. 52, no. 5, pp. 1230–1240, 2009.

[110] O. Schwartz, S. Gannot, and E. A. P. Habets, "An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1491–1506, Sept 2016.

[111] E. O. Selkirk, *Phonology and syntax: the relationship between sound and structure*. MIT press, 1986.

[112] K. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001.

[113] Q. Summerfield, "Lipreading and audio-visual speech perception," *Philosophical Transactions: Biological Sciences*, vol. 335, no. 1273, pp. 71–78, 1992.

[114] C. Taal *et al.*, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[115] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 7, pp. 1420–1434, 2009.

[116] P. Taylor, *Text-to-speech synthesis*. Cambridge university press, 2009.

[117] T. Van den Bogaert *et al.*, "Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids," *J. Acoust. Soc. Am.*, vol. 125, no. 1, pp. 360–371, 2009.

[118] H. L. Van Trees, "Arrays and spatial filters," in *Optimum Array Processing*, ser. Detection, Estimation, and Modulation Theory. Wiley, 2004, ch. 2, pp. 17–79.

[119] ——, *Optimum Array Processing*, ser. Detection, Estimation, and Modulation Theory. Wiley, 2004.

[120] T. van Waterschoot and M. Moonen, "Fifty years of acoustic feedback control: State of the art and future challenges," *Proc. IEEE*, vol. 99, no. 2, pp. 288–327, Feb 2011.

[121] K. Wagener, T. Brand, and B. Kollmeier, "Development and evaluation of a German sentence test, parts I–III," *Zeitschrift Fur Audiologie*, vol. 38, pp. 4–15, 44–56, and 86–95, 1999.

[122] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *Int. J. Audiology*, vol. 42, no. 1, pp. 10–17, 2003.

[123] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "A precedence effect in sound localization," *J. Acoust. Soc. Am.*, vol. 21, no. 4, pp. 468–468, 1949.

[124] Y. Xu *et al.*, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan 2014.

[125] H. Ye and R. D. DeGroat, "Maximum likelihood DOA estimation and asymptotic Cramér-Rao bounds for additive unknown colored noise," *IEEE Trans. Signal Process.*, vol. 43, no. 4, pp. 938–949, 1995.

[126] A. A. Zekveld, S. E. Kramer, and J. M. Festen, "Pupil response as an indication of effortful listening: The influence of sentence intelligibility," *Ear and Hearing*, vol. 31, no. 4, pp. 480–490, 2010.

[127] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 5, Apr 1988, pp. 2578–2581.

[128] W. Zhang, E. A. Habets, and P. A. Naylor, "On the use of channel shortening in multichannel acoustic system equalization," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, 2010.

Part II

Papers

Paper A

Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids

A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen

The paper has been presented at the
22nd European Signal Processing Conference (EUSIPCO), pp. 61–65,
Lisbon, Portugal 2014.

The layout has been revised.

Abstract

We propose a multi-channel Wiener filter for speech dereverberation in hearing aids. The proposed algorithm uses joint maximum likelihood estimation of the speech and late reverberation spectral variances, under the assumption that the late reverberant sound field is cylindrically isotropic. The dereverberation performance of the algorithm is evaluated using computer simulations with realistic hearing aid microphone signals including head-related effects. The algorithm is shown to work well with signals reverberated both by synthetic and by measured room impulse responses, achieving improvements in the order of 0.5 PESQ points and 5 dB frequency-weighted segmental SNR.

1 Introduction

Hearing impaired listeners experience increased difficulty in understanding speech in reverberant and noisy conditions [1]. In order to enable them to attain the same speech intelligibility as normal hearing persons, various signal enhancement algorithms are used in Hearing Aids (HAs). Both single- and multi-microphone (spatial) methods are commonly used in HAs, notably spectral modification and beamforming [2].

The Multi-channel Wiener Filter (MWF) [3] is a method which currently receives a lot of attention in the research community, e.g. [4], [5], [6]. Implementation of the MWF requires knowledge of the inter-microphone covariance matrices of the target signal (i.e. speech) and of the interference (e.g. ambient noise or reverberation). Traditionally a Voice Activity Detector (VAD) is used to enable noise covariance matrix estimation during speech pauses, e.g. [6]. This approach is based on the assumption that the interference covariance matrix is constant during speech presence. In reverberant conditions this assumption is not valid, which necessitates on-line estimation of the reverberation covariance matrix.

In the present study, we propose an MWF algorithm for speech dereverberation, which jointly estimates the target and interference spectral variances also during speech presence. The algorithm uses a Maximum Likelihood Estimation (MLE) method presented first in [7] which is novel in the speech dereverberation context. We assume a cylindrically isotropic spatial distribution of the late reverberation and a known speaker direction. Therefore, the structure of the inter-microphone covariance matrices of the speech and reverberation is known and only the time-varying spectral variances (the scaling factors of these matrices) are estimated in the MLE framework.

The proposed algorithm bears some similarities to the one presented in [4]. In both methods an isotropic spatial distribution of the late reverberant field is assumed and the spectral variances of the interference are estimated regardless of speech presence. However, while [4] uses intermediate "reference signals" (based on [5]) to estimate the reverberation variances, we compute

these estimates directly form the input covariance matrix (based on [7]). The method presented here is designed for and evaluated in a hearing aid usage scenario and with real room impulse responses, whereas in [4], microphones were assumed to reside in free field and reverberation was simulated using an image model of a rectangular room.

2 Signal model and assumptions

The proposed algorithm operates on M microphone signals represented as complex-valued Short Time Fourier Transform (STFT) coefficients. They are collected in a vector

$$\mathbf{y}(k,n) = [y_1(k,n) \dots y_m(k,n) \dots y_M(k,n)]^T, \tag{A.1}$$

where $y_m(k,n)$ is the STFT coefficient of the m-th microphone signal in the k-th frequency sub-band and the n-th time frame. Based on the assumption of signal independence between sub-bands, we will operate on them separately. This allows us to omit the frequency index k in the following description without loss of generality.

The input signal $\mathbf{y}(n)$ is assumed to be the sum of the target speech component $\mathbf{s}(n)$ and an interference component $\mathbf{v}(n)$. Both $\mathbf{s}(n)$ and $\mathbf{v}(n)$ are defined similarly to (A.1). The interference $\mathbf{v}(n)$ is assumed to be late reverberation, ambient noise, or a sum of both. In either case, it is assumed to be uncorrelated to the target speech component $\mathbf{s}(n)$. This allows us to model the covariance matrix of the input as the sum of the covariance matrices of the two signal components:

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(n) &= E\{\mathbf{y}(n)\mathbf{y}^H(n)\} \\ &= E\{\mathbf{s}(n)\mathbf{s}^H(n)\} + E\{\mathbf{v}(n)\mathbf{v}^H(n)\} \\ &= \mathbf{\Phi}_{\mathbf{s}}(n) + \mathbf{\Phi}_{\mathbf{v}}(n).\end{aligned} \tag{A.2}$$

We model the speaker as a point source and therefore the speech component can be expressed as

$$\mathbf{s}(n) = s(n)\mathbf{d}. \tag{A.3}$$

The scalar signal $s(n)$ represents the speech signal at a certain reference position, commonly chosen as one of the microphones. Elements of the vector $\mathbf{d}$ represent relative transfer functions of the speech signal between the reference position and all microphones of the array. The vector $\mathbf{d}$ is assumed to be known, and depends primarily on the microphone array geometry and on the direction of the speech source. In the beamforming context, we will refer to $\mathbf{d}$ as a steering vector.

We employ an isotropic model of the interference $\mathbf{v}(n)$. Taking this and (A.3) into account, (A.2) can be rewritten as

$$\mathbf{\Phi}_{\mathbf{y}}(n) = \underbrace{\phi_s(n)\mathbf{d}\mathbf{d}^H}_{\mathbf{\Phi}_{\mathbf{s}}(n)} + \underbrace{\phi_v(n)\mathbf{\Gamma}_{\text{iso}}}_{\mathbf{\Phi}_{\mathbf{v}}(n)}, \tag{A.4}$$

where $\phi_s(n)$ and $\phi_v(n)$ are, respectively, (scalar) spectral variances of the speech and of the interference component of the reference microphone signal. Because, in general, the speech and noise processes are non-stationary, their variances $\phi_s(n)$ and $\phi_v(n)$ are time-varying. The matrix $\mathbf{\Gamma}_{\text{iso}}$ is the normalized covariance matrix of the isotropic sound field, and similarly to $\mathbf{d}$, is assumed to be known and constant.

2.1 Discussion of validity of assumptions

The intended application of the proposed algorithm is intelligibility improvement of reverberant and/or noisy speech in HAs. Assumptions with regard to the employed signal model are made to capture aspects of the actual physical signals which are most relevant to this particular task and application.

In reverberant conditions, speech intelligibility is affected primarily by late reverberation, whereas early reflections are believed to be beneficial [8]. For that reason, the model of the interference was chosen to describe properties of specifically the late part of the reverberation.

In [9], the spatial energy distribution of reverberant sound fields was studied. It was shown that all spatial directions are represented in the late reverberant energy, but only few directions are in the energy of early reflections. This supports our assumption that the late reverberant field is isotropic.

An isotropic model of the ambient noise is also ecologically justified, especially in applications where there is no prior knowledge on the spatial distribution of the noise, e.g. in hearing aids. The spatial probability distribution of the noise impinging on the microphone array can reasonably be assumed uniform, i.e. isotropic.

The assumption of $\mathbf{d}$ being known is reasonable in hearing aid design. It is supported by the fact that, in most situations, the hearing aid user is looking at the person he is speaking with (e.g. to facilitate lip reading). Hence, $\mathbf{d}$ corresponds to a target source frontal to the HA user.

In the present work, the interference $\mathbf{v}(n)$ is modeled as independent, and therefore uncorrelated with the speech signal $\mathbf{s}(n)$. This assumption is natural for the ambient noise but is questionable with regard to the reverberation. Our rationale is that the late part of the room impulse responses is considerably disrupted by thermal fluctuations [10] and small movements of the source and microphone array [11]. These instabilities are unavoidable in real use of a HA and effectively decorrelate the late reverberation from the direct sound.

3 Multi-channel Wiener filter

It is well known that the MWF is the Linear Minimum Mean Square Error (LMMSE) solution to the problem of signal estimation in a setup presented in Section 2, [3]. It is also well known that the MWF can be factorized into a

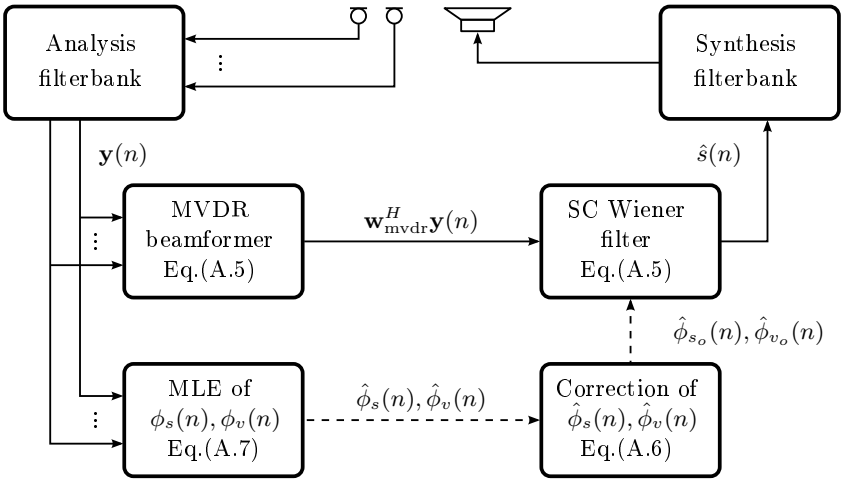


Fig. A.1: Block diagram of the proposed algorithm.

Minimum Variance Distortionless Response (MVDR) beamformer and a Single Channel (SC) Wiener filter [3].

The structure of the proposed MWF-type algorithm is depicted in Fig. A.1. The signal resulting from the MWF is the LMMSE estimate of the target speech signal at the reference position and may be written as

$$\hat{s}(n) = \mathbf{w}_{\text{mwf}}^H(n)\mathbf{y}(n), \text{ where} \tag{A.5a}$$

$$\mathbf{w}_{\text{mwf}}(n) = \underbrace{\left[\frac{\phi_{s_o}(n)}{\phi_{s_o}(n)+\phi_{v_o}(n)}\right]}_{g_{\text{sc}}(n)} \underbrace{\frac{\mathbf{\Gamma}_{\text{iso}}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{\Gamma}_{\text{iso}}^{-1}\mathbf{d}}}_{\mathbf{w}_{\text{mvdr}}}. \tag{A.5b}$$

In (A.5a–A.5b) the vector of MVDR beamformer coefficients and the SC Wiener filter gain have been denoted as $\mathbf{w}_{\text{mvdr}}$ and $g_{\text{sc}}(n)$, respectively. $\phi_{s_o}(n)$ and $\phi_{v_o}(n)$ denote the spectral variances of the speech and the interference at the output of the MVDR beamformer. They can be expressed as

$$\phi_{s_o}(n) = \phi_s(n), \tag{A.6a}$$

$$\phi_{v_o}(n) = \phi_v(n)(\mathbf{d}^H\mathbf{\Gamma}_{\text{iso}}^{-1}\mathbf{d})^{-1}. \tag{A.6b}$$

The MVDR beamformer does not distort the variance of the speech (A.6a), but the variance of the interference has to be corrected by the beamformer suppression factor (A.6b) [3]. It is important to note that $\mathbf{w}_{\text{mvdr}}$ depends only on $\mathbf{\Gamma}_{\text{iso}}$ and $\mathbf{d}$. Because we assume that these are known and constant, the beamformer coefficients $\mathbf{w}_{\text{mvdr}}$ can be calculated beforehand.

The SC Wiener filter gain $g_{\text{sc}}(n)$ is time-varying and depends on the spectral variances $\phi_s(n)$ and $\phi_v(n)$. They are unknown and have to be estimated from the noisy/reverberant observations $\mathbf{y}(n)$ for each time frame and frequency bin.

Several methods exist for estimating $\phi_s(n)$ and $\phi_v(n)$, e.g. [4], [5], [7]. The proposed algorithm uses MLEs which were derived by Ye and DeGroat [7] for a similar signal model to the one employed in the present study, although in a non-acoustic context. These MLEs may be expressed as

$$\hat{\phi}_v(n) = \frac{1}{M-1} \operatorname{tr}\left\{ \left(\mathbf{I} - \mathbf{d}\mathbf{w}_{\mathrm{mvdr}}^H \right) \hat{\mathbf{\Phi}}_{\mathbf{y}}(n)\, \mathbf{\Gamma}_{\mathrm{iso}}^{-1} \right\}, \tag{A.7a}$$

$$\hat{\phi}_s(n) = \mathbf{w}_{\mathrm{mvdr}}^H \left(\hat{\mathbf{\Phi}}_{\mathbf{y}}(n) - \hat{\phi}_v(n)\, \mathbf{\Gamma}_{\mathrm{iso}} \right) \mathbf{w}_{\mathrm{mvdr}}, \tag{A.7b}$$

where $\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)$ is the estimate of the covariance matrix of the input signal, and $\mathrm{tr}\{\cdot\}$ denotes the matrix trace operator.

4 Experimental setup

In order to evaluate the performance of the proposed algorithm, a series of computer simulations was conducted. Technical details on these simulations are described in Sections 4.1–4.2 and the evaluation results are discussed in Section 5

4.1 Speech signals and room impulse responses

Recorded speech utterances of male and female native English speakers were obtained from the TIMIT database [12]. Individual utterances were concatenated into longer sequences and artificially reverberated by convolving them with either synthetic or measured multi-channel Room Impulse Responses (RIRs). Each RIR consisted of 4 channels corresponding to the microphones of a pair of 2-microphone Oticon Epoq Behind-The-Ear (BTE) HAs placed on the ears of a Brüel&Kjær Head And Torso Simulator (HATS).

Five RIRs were recorded in real rooms with the source of the probe sound placed in front of the HATS at a distance between 0.9 m and 2 m. The reverberation time T_{60}, the clarity index C_{50} and the Direct-to-Reverberation Ratio (DRR) calculated from these RIRs are given in the upper part of Table A.1. In none of the used rooms the reverberation was truly isotropic and in some of them it was strongly dominated by certain directions (especially in the bathroom and auditorium). In that sense, the used RIRs constitute a fair sample of reverberant conditions a hearing aid user might encounter.

A sixth, synthetic RIR was designed to measure the performance of the proposed algorithm in conditions completely matching the underlying assumptions on reverberation isotropy. The reverberation tail of the synthetic RIR was modeled by a sum of 72 exponentially decaying independent white noise sequences, filtered through anechoic Head Related Transfer Functions (HRTFs) measured for 72 evenly spaced positions on the horizontal circle of the HATS. The direct path component of this RIR was computed from the HRTF of a frontally placed sound source. The HRTFs were recorded with an equivalent

Table A.1: Acoustic parameters of the rooms simulated in the evaluation experiment.

Room	T_{60} [s]	C_{50} [dB]	DRR [dB]
Bathroom	0.8	5.2	−10.1
Cellar	1.2	5.7	2.2
Staircase	2.3	11.0	4.1
Office	1.4	8.8	2.3
Auditorium	1.3	13.4	5.2
Isotropic	1.0	4.7	−0.4

HA/HATS combination as the real RIR measurements. Parameters of the synthesized RIR are given in the last row of Table A.1 (denoted as "Isotropic").

4.2 Implementation of the proposed algorithm

The simulated reverberant microphone signals were transformed into time-frequency samples $y_m(k,n)$ using an STFT filterbank. An inverse STFT combined with an overlap-add procedure was used to resynthesize the output signal (see Fig. A.1). The frame length of the analysis was 8 ms with 50% overlap between consecutive frames. Traditionally, longer frame lengths are used in speech processing, however, in hearing aids short processing delay is a strong design constraint. A square root Hann window function was used in both the analysis and the synthesis filterbank. A sampling frequency of 16 kHz was used based on the assumption that frequencies above 8 kHz are negligible in speech perception.

In order to implement the algorithm with (A.5), (A.6), and (A.7), $\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)$, $\mathbf{d}$, and $\mathbf{\Gamma}_{\text{iso}}$ are needed. The input covariance matrix $\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)$ was estimated from $\mathbf{y}(n)$ using recursive averaging with a time constant of 40 ms.

For each reverberant condition a different steering vector $\mathbf{d}$ was calculated from the respective RIR truncated to the part containing only the direct path response. Vectors $\mathbf{d}$ were computed by discrete Fourier transformation of the truncated RIRs after appropriate zero-padding. In the synthetic reverberation condition, $\mathbf{d}$ was computed from the anechoic impulse response of the target direction.

The normalized covariance matrix of the isotropic sound field $\mathbf{\Gamma}_{\text{iso}}$ was modeled as

$$\mathbf{\Gamma}_{\text{iso}} = \frac{1}{S}\sum_{s=1}^{S} \mathbf{d}_{\text{hrtf}}(\alpha_s)\mathbf{d}_{\text{hrtf}}^{H}(\alpha_s), \tag{A.8}$$

where each relative transfer function vector $\mathbf{d}_{\text{hrtf}}(\alpha_s)$ corresponded to the HRTF measured in an anechoic chamber for one of the azimuth angles $\alpha_s \in \{5°, 10°, \ldots, 360°\}$ using the HA/HATS. In this way, $\mathbf{\Gamma}_{\text{iso}}$ represents

the frequency-dependent inter-microphone covariance matrix (up to a scalar multiplication) of a cylindrically isotropic sound field.

5 Performance evaluation

The evaluation of the proposed algorithm was based on three objective performance measures: Speech-to-Reverberation Modulation energy Ratio (SRMR) [13], Frequency-Weighted Segmental SNR (FWSegSNR) [14] and Perceptual Evaluation of Speech Quality (PESQ) [14]. Their Matlab implementations were obtained from the 2014 Reverb Challenge [15] website. The evaluation results are presented in Fig. A.2.

The three performance measures were calculated for: the unprocessed reverberant signal $y_1(n)$, the signal processed by the beamformer only ($\mathbf{w}_{\text{mvdr}}^H \mathbf{y}(n)$), and the reverberant signal enhanced by the full algorithm ($\hat{s}(n)$) (see (A.5) and Fig. A.1). The results calculated from these signals are denoted as "Input", "MVDR", and "MWF", respectively. The proposed algorithm was evaluated for two different microphone array configurations: 4-microphone (using both HAs), and 2-microphone (using only the left HA). In the 4-microphone case we assume that the signals are communicated between the two hearing aids instantly and without error.

The reference signal used to compute FWSegSNR and PESQ was the direct path speech signal $s(n)$. In case of the SRMR, which is a non-intrusive measure, the score of the reference signal was also computed and is presented in Fig. A.2(b).

5.1 Discussion of results

For the simulations with synthetic isotropic reverberation (denoted as "Isotropic"), the proposed algorithm results in an increase of all considered performance measures. Both the MVDR beamformer and the SC Wiener filter stages of the algorithm contribute positively to that increase. Moreover, the 4-microphone configuration results in a better performance than the 2-microphone configuration. This is an indication, that the proposed method is able to use and benefit from the additional spatial information available in the 4-microphone setup, i.e. when two HAs are used.

In most simulations with RIRs measured in real rooms the increase in the performance measures was lower than in the synthetic isotropic reverberation condition. Nonetheless, in some cases the improvement was of similar magnitude (in the cellar, staircase, and office conditions). This suggests that the isotropic late reverberation model is sufficiently accurate in many real-world reverberant environments and can be used to effectively dereverberate the signal. The increase of the performance scores was smaller in simulations using the RIR of the auditorium, and even negative in the bathroom condition (PESQ and FWSegSNR). Analysis of these two RIRs revealed that the isotropy assump-

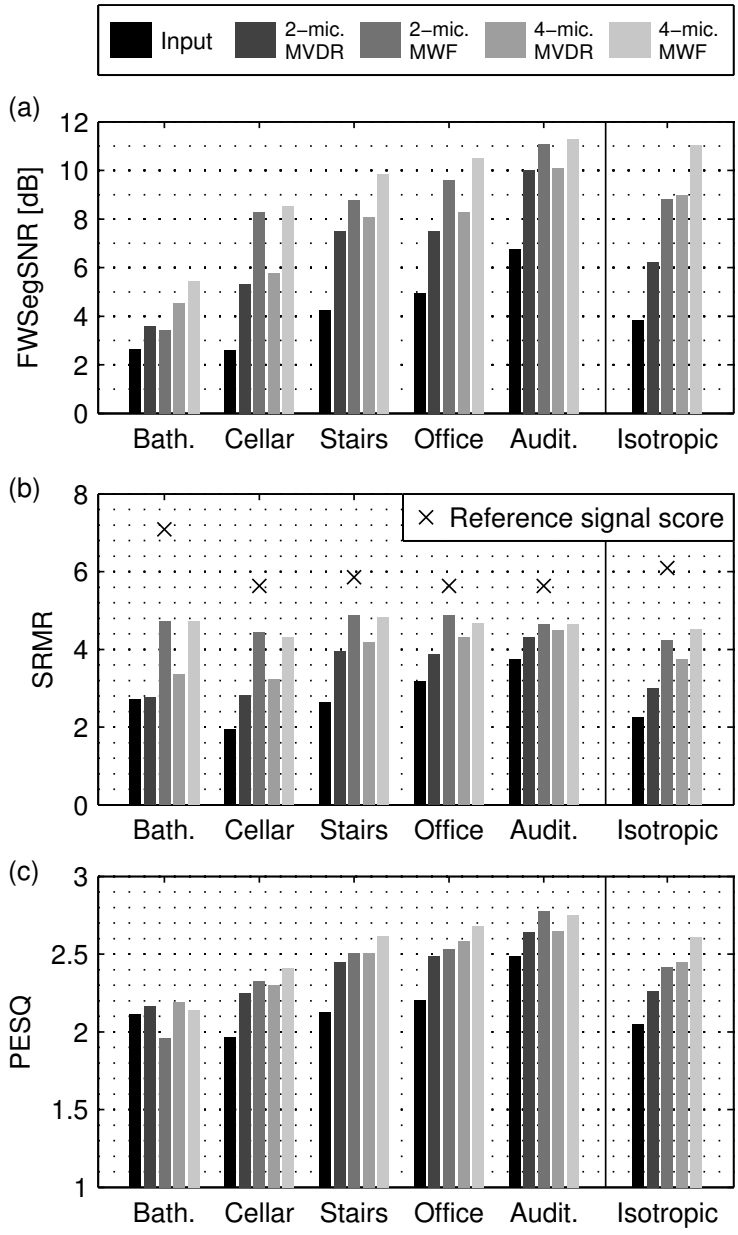


Fig. A.2: (a) FWSegSNR, (b) SRMR and (c) PESQ scores of the reverberant ("Input"), and processed ("MVDR" and "MWF") signals for different reverberation conditions, and configurations of the microphone array.

tion was not valid in these situations because of isolated specular reflections dominating the reverberation.

The sound quality and speech intelligibility of the processed signals was subjectively assessed through informal listening tests. The perceptual gain from using the algorithm was most pronounced in the simulated isotropic reverberation condition. In the cellar, staircase and the office conditions, the speech was audibly dereverberated and the sound quality was almost unaffected. In the auditorium and particularly in the bathroom conditions, sound artifacts were noticeable.

It is relevant to mention, that the algorithm proposed in this paper is also applicable to target signals other than speech and to interference types other than reverberation. However, it is a prerequisite that the spatial distribution of the interference is isotropic or is otherwise known or estimated. Although the evaluation of the proposed algorithm was conducted in reverberant-only condition, it is reasonable to expect similar performance in an arbitrary isotropic non-stationary noise.

6 Conclusion

In this paper we have proposed a Multi-channel Wiener Filter (MWF) which uses joint Maximum Likelihood Estimation (MLE) of speech and reverberation spectral variances. The MLE method was adopted from the work of Ye and DeGroat [7]. The proposed MWF algorithm was implemented and its speech dereverberation performance for hearing aids was evaluated. It was shown that the proposed algorithm performs well in both synthetic and realistic reverberation conditions. The performance of the proposed method was best when the assumption on the interference isotropy was close to valid. In non-isotropic reverberation/ambient noise conditions on-line estimation of the interference covariance matrix structure could be used to improve the performance. This is a topic for future research.

References

[1] A. K. Nabelek and J. Pickett, "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech Hearing Res.*, vol. 17, no. 4, pp. 724–739, 1974.

[2] V. Hamacher *et al.*, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 2915–2929, Jan. 2005.

[3] K. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001.

[4] S. Braun and E. A. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. 21st Eur. Signal Process. Conf. (EUSIPCO)*, Marrakech, Morocco, 2013, pp. 1–5.

[5] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Bucharest, Romania, 2012, pp. 295–299.

[6] B. Cornelis, M. Moonen, and J. Wouters, "Speech intelligibility improvements with hearing aids using bilateral and binaural adaptive multichannel wiener filtering based noise reduction," *J. Acoust. Soc. Am.*, vol. 131, no. 6, pp. 4743–4755, 2012.

[7] H. Ye and R. D. DeGroat, "Maximum likelihood DOA estimation and asymptotic Cramér-Rao bounds for additive unknown colored noise," *IEEE Trans. Signal Process.*, vol. 43, no. 4, pp. 938–949, 1995.

[8] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am.*, vol. 113, no. 6, pp. 3233–3244, 2003.

[9] B. N. Gover, J. G. Ryan, and M. R. Stinson, "Measurements of directional properties of reverberant sound fields in rooms using a spherical microphone array," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2138–2148, 2004.

[10] G. W. Elko, E. Diethorn, and T. Gänsler, "Room impulse response variation due to temperature fluctuations and its impact on acoustic echo cancellation," in *Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Kyoto, Japan, 2003, pp. 67–70.

[11] J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *J. Sound and Vibration*, vol. 102, no. 2, pp. 217–228, 1985.

[12] J. S. Garofolo *et al.*, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM.* NIST, 1993.

[13] T. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[14] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan 2008.

[15] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The reverb challenge: Acommon evaluation framework for dereverberation and recognition of reverberant speech," in *IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, Oct 2013, pp. 1–4.

Paper B

Multi-channel PSD estimators for speech dereverberation – a theoretical and experimental comparison

A. Kuklasiński, S. Doclo, T. Gerkmann, S. H. Jensen,
and J. Jensen

The paper has been presented at the
IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 91–95, Brisbane, Australia, 2015.

The layout has been revised.

Abstract

In this paper we perform an extensive theoretical and experimental comparison of two recently proposed multi-channel speech dereverberation algorithms. Both of them are based on the multi-channel Wiener filter but they use different estimators of the speech and reverberation power spectral densities (PSDs). We first derive closed-form expressions for the mean square error (MSE) of both PSD estimators and then show that one estimator – previously used for speech dereverberation by the authors – always yields a better MSE. Only in the case of a two microphone array or for special spatial distributions of the interference both estimators yield the same MSE. The theoretically derived MSE values are in good agreement with numerical simulation results and with instrumental speech quality measures in a realistic speech dereverberation task for binaural hearing aids.

1 Introduction

Background noise and reverberation may have a detrimental effect on speech quality and intelligibility [1]. Consequently, speech denoising and dereverberation algorithms are of interest in many applications, e.g. hearing aids, mobile phones, etc. Many of these devices contain multiple microphones, which enables the use of spatial filtering algorithms such as the Multi-channel Wiener Filter (MWF) [2, 3]. Under a set of commonly made assumptions the MWF is an optimal estimator of the speech signal in the Minimum Mean Square Error (MMSE) sense [2]. However, in order to obtain its theoretical performance the MWF requires knowledge of the (cross-) Power Spectral Density (PSD) matrices of the target (speech) and interference (noise, reverberation) signal components. These are usually unknown and have to be estimated from the noisy and reverberant microphone signals. In practice, the performance of the resulting MWF depends on the accuracy of the used PSD estimation scheme.

In this paper we compare two multi-channel speech dereverberation algorithms recently proposed in [4] and [5]. Both algorithms are based on the MWF and use the assumption that the reverberant sound field is cylindrically isotropic. The PSD estimators used in [4] and [5] are both derived using the Maximum (ML) methodology but use different statistical assumptions, and therefore yield different formulas and results.

In order to perform a theoretical comparison of the two PSD estimation schemes we first derive analytical expressions for their Mean Square Error (MSE). This allows us to show that the PSD estimators used in [4] achieve lower or equal MSE compared to the PSD estimators in [5]. We also derive the conditions under which the two PSD estimation schemes yield the same MSE. We verify these theoretical results using numerical simulations.

Finally, we evaluate the speech dereverberation performance of the MWFs from [4] and [5] in a simulation of binaural hearing aids in realistic rever-

berant conditions. The results of the experiment show that the algorithm from [4] outperforms [5] in terms of objective performance measures such as Frequency-Weighted Segmental SNR (FWSegSNR) [6] and Perceptual Evaluation of Speech Quality (PESQ) [7].

2 Signal model and assumptions

The signal model and assumptions that are used in the speech dereverberation algorithms proposed in [4] and [5] share many characteristics. Both algorithms operate on Short Time Fourier Transform (STFT) coefficients $y_m(k,n)$ which are computed from the time domain signals $y_m(t)$ of M microphones:

$$y_m(k,n) = \sum_{t=0}^{T-1} y_m(t+nD)w(t)e^{-2\pi i k\frac{t}{T}}, \qquad m = 1,\dots,M,$$

where n is the time frame index and k is the frequency bin index. The STFT order is denoted by T, the filterbank decimation factor by D, and $w(t)$ is the analysis window function. The algorithms from [4] and [5] process the individual frequency bins independently of each other. This enables us to omit the index k without loss of generality. For notational convenience the STFT coefficients corresponding to all microphones are stacked in a vector as: $\mathbf{y}(n) = [\,y_1(n)\dots y_M(n)\,]^T$.

The algorithms from [4] and [5] employ an additive model of the reverberant speech signal:

$$\mathbf{y}(n) = \mathbf{s}(n) + \mathbf{v}(n) \overbrace{+\mathbf{x}(n)}^{\text{only in [5]}}, \tag{B.1}$$

where $\mathbf{s}(n)$ denotes the direct-path speech component and $\mathbf{v}(n)$ denotes the reverberation component of the microphone signal. The algorithm from [5] allows for an additional noise term $\mathbf{x}(n)$, whose cross-PSD matrix must be known. In this study, for mathematical convenience, we assume that this additional noise component is equal to zero. This corresponds to an assumption that $\mathbf{x}(n)$ is negligible compared to the reverberation component, which may be valid in some situations. It is assumed that the vectors $\mathbf{s}(n)$ and $\mathbf{v}(n)$ are statistically independent across time frames and frequency bins.

The algorithms from [4] and [5] aim to estimate the direct-path speech signal component $s(n)$ at a certain reference position, e.g. one of the microphones. Because the speech is assumed to be generated by a point source, the vector $\mathbf{s}(n)$ may be written as the product of $s(n)$ and a vector of Relative Transfer Functions (RTFs) $\mathbf{d}$ [8]:

$$\mathbf{y}(n) = s(n)\mathbf{d} + \mathbf{v}(n).$$

The elements of $\mathbf{d}$ correspond to the transfer functions of the direct-path speech from the chosen reference position to all microphones. In [4] and [5] the RTF vector $\mathbf{d}$ is assumed to be known.

In both algorithms the focus is on reducing the late part of the reverberation, which is assumed to be uncorrelated with the direct-path speech. Hence, the cross-PSD matrix of $\mathbf{y}(n)$ can be modeled as the sum of the speech and the reverberation cross-PSD matrices:

$$\mathbf{\Phi}_{\mathbf{y}}(n) = E\big[\mathbf{y}(n)\mathbf{y}^H(n)\big] = \mathbf{\Phi}_{\mathbf{s}}(n) + \mathbf{\Phi}_{\mathbf{v}}(n).$$

where $E[\,]$ denotes the expectation operator. Because of the assumption that the speech is generated by a point source, $\mathbf{\Phi}_{\mathbf{s}}(n)$ is modeled as a rank-one matrix and can be written in terms of the scalar PSD $\phi_s(n)$ of the direct-path speech at the reference position and the RTF vector $\mathbf{d}$: $\phi_s(n)\mathbf{d}\mathbf{d}^H$. Similarly, matrix $\mathbf{\Phi}_{\mathbf{v}}(n)$ may be written as a product of the scalar PSD $\phi_v(n)$ of the reverberation at the reference position, and the cross-PSD matrix $\mathbf{\Gamma}_{\mathbf{v}}$ of the reverberation normalized by $\phi_v(n)$:

$$\mathbf{\Phi}_{\mathbf{y}}(n) = \phi_s(n)\mathbf{d}\mathbf{d}^H + \phi_v(n)\mathbf{\Gamma}_{\mathbf{v}}, \tag{B.2}$$

Due to the assumption of cylindrical isotropy of the reverberant sound field made in both [4] and [5], the matrix $\mathbf{\Gamma}_{\mathbf{v}}$ is assumed to be constant, full-rank, and known. For free-field microphone arrays, $\mathbf{\Gamma}_{\mathbf{v}}$ can even be calculated analytically using information on microphone array geometry (as in [5]). Alternatively, e.g. for hearing aid applications, $\mathbf{\Gamma}_{\mathbf{v}}$ may be estimated from measurements using the actual microphone array in a (possibly simulated) isotropic sound field (as in [4]). While the vector $\mathbf{d}$ and the matrix $\mathbf{\Gamma}_{\mathbf{v}}$ are assumed to be known and constant, the PSDs $\phi_s(n)$ and $\phi_v(n)$ are unknown and time-varying because of the non-stationarity of $\mathbf{s}(n)$ and $\mathbf{v}(n)$.

3 Multi-channel Wiener filter

The algorithms from [4] and [5] are both based on the Multi-channel Wiener Filter (MWF) [2, 3]. It is well-known that the MWF is an MMSE-optimal estimator of the target speech $s(n)$ if the input signal components $\mathbf{s}(n)$ and $\mathbf{v}(n)$ are normally distributed, or alternatively, if the search is limited to linear estimators. Because of the rank-one assumption on $\mathbf{\Phi}_{\mathbf{s}}(n)$, the MWF may be factorized into an MVDR beamformer $\mathbf{w}_{\text{mvdr}}$ and a single-channel Wiener filter $g_{\text{sc}}(n)$ [2]:

$$\begin{aligned} \hat{s}(n) &= \mathbf{w}_{\text{mwf}}^H(n)\mathbf{y}(n), \\ \mathbf{w}_{\text{mwf}}(n) &= \underbrace{\left[\frac{\phi_{s_o}(n)}{\phi_{s_o}(n) + \phi_{v_o}(n)}\right]}_{g_{\text{sc}}(n)} \underbrace{\frac{\mathbf{\Gamma}_{\mathbf{v}}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{\Gamma}_{\mathbf{v}}^{-1}\mathbf{d}}}_{\mathbf{w}_{\text{mvdr}}}, \end{aligned} \tag{B.3}$$

where $\phi_{s_o}(n)$ and $\phi_{v_o}(n)$ are the PSDs of the direct-path speech and reverberation at the output of the MVDR beamformer, i.e.: $\phi_{s_o}(n) = \phi_s(n)$, and $\phi_{v_o}(n) = \mathbf{w}_{\text{mvdr}}^H\phi_v(n)\mathbf{\Gamma}_{\mathbf{v}}\mathbf{w}_{\text{mvdr}}$. For the signal model described in Sec. 2 the vector $\mathbf{w}_{\text{mvdr}}$ is constant and is readily calculated from $\mathbf{d}$ and $\mathbf{\Gamma}_{\mathbf{v}}$.

4 Power spectral density estimation

The main difference between the algorithms from [4] and [5] is the method used to estimate the unknown PSDs of the direct-path speech $\phi_s(n)$ and of the reverberation $\phi_v(n)$. In this section, we briefly review these two PSD estimation schemes.

4.1 Algorithm [4] by Kuklasiński et al.

The PSD estimators used in [4] are based on the assumption that the STFT coefficients of the signal components are circularly-symmetric complex Gaussian distributed, i.e.:

$$\mathbf{s}(n) \sim \mathcal{CN}\big(\mathbf{0}, \mathbf{\Phi}_{\mathbf{s}}(n)\big), \qquad \mathbf{v}(n) \sim \mathcal{CN}\big(\mathbf{0}, \mathbf{\Phi}_{\mathbf{v}}(n)\big).$$

The above distributions can be used to construct a likelihood function, compute its partial derivatives, and ultimately, derive a pair of joint Maximum Likelihood Estimators (MLEs) of $\phi_s(n)$ and $\phi_v(n)$. Several formulations of these estimators are available in the literature [9, 10], but in [4] the one from [9] has been used:

$$\hat{\phi}_{v,\,[4]}(n) = \frac{1}{M-1}\,\mathrm{tr}\Big[\big(\mathbf{I} - \mathbf{d}\mathbf{w}_{\mathrm{mvdr}}^{H}\big)\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)\;\mathbf{\Gamma}_{\mathbf{v}}^{-1}\Big], \tag{B.4a}$$

$$\hat{\phi}_{s,\,[4]}(n) = \mathbf{w}_{\mathrm{mvdr}}^{H}\Big[\hat{\mathbf{\Phi}}_{\mathbf{y}}(n) - \hat{\phi}_{v,\,[4]}(n)\;\mathbf{\Gamma}_{\mathbf{v}}\Big]\mathbf{w}_{\mathrm{mvdr}}, \tag{B.4b}$$

where tr[] denotes the matrix trace, $\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)$ denotes the estimate of the cross-PSD matrix of the input signal:

$$\hat{\mathbf{\Phi}}_{\mathbf{y}}(n) = \frac{1}{L}\sum_{l=0}^{L-1}\mathbf{y}(n-l)\mathbf{y}^{H}(n-l), \tag{B.5}$$

and where the PSDs $\phi_s(n)$ and $\phi_v(n)$ are assumed to be constant across the L averaged STFT frames.

4.2 Algorithm [5] by Braun and Habets

Similarly to [4], in [5] the reverberation PSD estimator is derived using the ML methodology. However, the likelihood function used in the derivation is based on a different statistical assumption than in [4], resulting in a different estimator.

Specifically, the reverberation PSD estimator used in [5] is derived by first defining a blocking matrix $\mathbf{B} \in \mathbb{C}_{M\times(M-1)}$ which represents a set of $M-1$ target-canceling beamformers. In [5] it is computed according to the method used in [10]:

$$\big[\mathbf{B}\;\mathbf{b}\big] = \mathbf{A}, \quad \mathbf{A} = \mathbf{I} - \mathbf{d}(\mathbf{d}^{H}\mathbf{d})^{-1}\mathbf{d}^{H}.$$

Next, an error matrix $\mathbf{\Phi}_{\text{err}}(n)$ is defined as:

$$\mathbf{\Phi}_{\text{err}}(n) = \hat{\tilde{\mathbf{\Phi}}}_{\mathbf{y}}(n) - \phi_v(n)\tilde{\mathbf{\Gamma}}_{\mathbf{v}}, \tag{B.6}$$

with $\tilde{\mathbf{\Gamma}}_{\mathbf{v}} = \mathbf{B}^H\mathbf{\Gamma}_{\mathbf{v}}\mathbf{B}$ and

$$\hat{\tilde{\mathbf{\Phi}}}_{\mathbf{y}}(n) = \mathbf{B}^H\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)\mathbf{B}. \tag{B.7}$$

The matrix $\hat{\tilde{\mathbf{\Phi}}}_{\mathbf{y}}(n)$ is the estimate of the cross-PSD matrix of the blocked input signal $\tilde{\mathbf{y}}(n) = \mathbf{B}^H\mathbf{y}(n)$. Because $\mathbf{B}^H\mathbf{s}(n) = \mathbf{0}$, $\hat{\tilde{\mathbf{\Phi}}}_{\mathbf{y}}(n)$ is equivalently the estimate of the cross-PSD matrix of the blocked reverberation signal component $\mathbf{B}^H\mathbf{v}(n)$ (cf. (B.1)). Hence, the matrix $\mathbf{\Phi}_{\text{err}}(n)$ in (B.6) can be interpreted as the error between the blocked reverberation cross-PSD matrix $\phi_v(n)\tilde{\mathbf{\Gamma}}_{\mathbf{v}}$ (cf. (B.2)) and its estimate $\hat{\tilde{\mathbf{\Phi}}}_{\mathbf{y}}(n)$. In [5] the elements of $\mathbf{\Phi}_{\text{err}}(n)$ are modeled as independent circularly-symmetric complex Gaussian random variables of equal variance. This assumption is used to construct a likelihood function from which an MLE of $\phi_v(n)$ is calculated as [5]:

$$\hat{\phi}_{v,\,[5]}(n) = \text{tr}\Big[\tilde{\mathbf{\Gamma}}_{\mathbf{v}}\hat{\tilde{\mathbf{\Phi}}}_{\mathbf{y}}(n)\Big]\text{tr}\Big[\tilde{\mathbf{\Gamma}}_{\mathbf{v}}^2\Big]^{-1}. \tag{B.8a}$$

The corresponding estimator of $\phi_s(n)$ is derived without the use of the ML methodology, but coincidentally has the same form as the MLE used in [4] (B.4b):

$$\hat{\phi}_{s,\,[5]}(n) = \mathbf{w}_{\text{mvdr}}^H\Big[\hat{\mathbf{\Phi}}_{\mathbf{y}}(n) - \hat{\phi}_{v,\,[5]}(n)\ \mathbf{\Gamma}_{\mathbf{v}}\Big]\mathbf{w}_{\text{mvdr}}. \tag{B.8b}$$

5 Analytical evaluation

In this section we analytically derive the MSE of the reverberation PSD estimator from [5] and compare it to the MSE of the reverberation PSD estimator from [4]. Differences between the direct-path speech PSD estimators from [4] and [5] are exclusively due to the different reverberation PSD estimators used in (B.4b) and (B.8b). Therefore, relations between the MSEs of the direct-path speech PSD estimators are completely determined by and are analogous to the relations between the MSEs of the reverberation PSD estimators.

We start by noting that the PSD estimators from [4] and [5] are unbiased (without proof):

$$E\big[\hat{\phi}_{p,r}(n)\big] = \phi_p(n), \quad p \in \{s, v\}, \quad r \in \{[4], [5]\}. \tag{B.9}$$

Hence, the MSEs of these estimators are identical to their variances.

The variance of the direct-path speech PSD estimator from [4] can be shown to be equal to the corresponding asymptotic Cramér-Rao Lower Bound (CRLB) which is equal to [11]:

$$\begin{aligned}\text{var}\big(\hat{\phi}_{s,\,[4]}(n)\big) &= \text{CRLB}\big(\hat{\phi}_s(n)\big) \\ &= \phi_s^2(n)\frac{1}{L}\left[\left(\frac{1+\xi(n)}{\xi(n)}\right)^2 + \frac{1}{M-1}\frac{1}{\xi^2(n)}\right],\end{aligned} \tag{B.10}$$

where $\xi(n) = \phi_{s_o}(n)/\phi_{v_o}(n)$ is the SNR at the output of the MVDR beamformer. From [11] it also follows that the variance of the reverberation PSD estimator from [4] is equal to the respective CRLB, and that this CRLB is equal to:

$$\mathrm{var}\big(\hat{\phi}_{v,\,[4]}(n)\big) = \mathrm{CRLB}\big(\hat{\phi}_v(n)\big) = \phi_v^2(n)\frac{1}{L}\frac{1}{M-1}. \tag{B.11}$$

We derive the variance of the reverberation PSD estimator from [5] by using (B.8a), (B.7) and (B.5) and moving the deterministic factors outside the variance operator:

$$\mathrm{var}\big(\hat{\phi}_{v,\,[5]}(n)\big) = \mathrm{tr}\Big[\tilde{\mathbf{\Gamma}}_{\mathbf{v}}^2\Big]^{-2}\frac{1}{L}\,\mathrm{var}\Big(\mathrm{tr}\Big[\tilde{\mathbf{y}}^H(n)\tilde{\mathbf{\Gamma}}_{\mathbf{v}}\tilde{\mathbf{y}}(n)\Big]\Big),$$

The trace operator inside the variance operator may now be omitted because its argument has been reduced to a quadratic form (a scalar). The variance of such quadratic forms in circularly-symmetric complex Gaussian random vectors is given by [12, p. 513, eq. (15.30)]:

$$\mathrm{var}\big(\mathbf{a}^H\mathbf{Z}\mathbf{a}\big) = \mathrm{tr}\big(\mathbf{\Phi}_{\mathbf{a}}\mathbf{Z}\mathbf{\Phi}_{\mathbf{a}}\mathbf{Z}\big),\ \text{where}\ \mathbf{a}\sim\mathcal{CN}\big(\mathbf{0},\mathbf{\Phi}_{\mathbf{a}}\big). \tag{B.12}$$

Using (B.12) and the fact that $\tilde{\mathbf{y}}(n)\sim\mathcal{CN}\big(\mathbf{0},\phi_v(n)\tilde{\mathbf{\Gamma}}_{\mathbf{v}}\big)$, we obtain:

$$\mathrm{var}\big(\hat{\phi}_{v,\,[5]}(n)\big) = \phi_v^2(n)\frac{1}{L}\,\mathrm{tr}\Big[\tilde{\mathbf{\Gamma}}_{\mathbf{v}}^4\Big]\,\mathrm{tr}\Big[\tilde{\mathbf{\Gamma}}_{\mathbf{v}}^2\Big]^{-2}. \tag{B.13}$$

Before comparing (B.11) and (B.13), we transform (B.13) into a more convenient form. Let $\tilde{\mathbf{\Gamma}}_{\mathbf{v}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H$ denote the eigenvalue decomposition of the positive-definite Hermitian matrix $\tilde{\mathbf{\Gamma}}_{\mathbf{v}}$, where $\mathbf{\Lambda}$ is a diagonal matrix containing the $M-1$ positive eigenvalues $\lambda_1,\ldots,\lambda_{M-1}$ of $\tilde{\mathbf{\Gamma}}_{\mathbf{v}}$. Using the facts that $\mathrm{tr}(\tilde{\mathbf{\Gamma}}_{\mathbf{v}}) = \sum_{m=1}^{M-1}\lambda_m$, $\tilde{\mathbf{\Gamma}}_{\mathbf{v}}^p = \mathbf{V}\mathbf{\Lambda}^p\mathbf{V}^H$, and defining $\gamma_m = \lambda_m^2$, (B.13) may be written as:

$$\mathrm{var}\big(\hat{\phi}_{v,\,[5]}(n)\big) = \phi_v^2(n)\frac{1}{L}\,\frac{\sum_{m=1}^{M-1}\gamma_m^2}{\big(\sum_{m=1}^{M-1}\gamma_m\big)^2}\,. \tag{B.14}$$

If we denote the average of the squared eigenvalues γ_m by $\bar{\gamma}$, and the sample variance of these squared eigenvalues around $\bar{\gamma}$ by $\tilde{\gamma}^2$,

$$\bar{\gamma} = \frac{1}{M-1}\sum_{m=1}^{M-1}\gamma_m, \quad \tilde{\gamma}^2 = \left(\frac{1}{M-1}\sum_{m=1}^{M-1}\gamma_m^2\right) - \bar{\gamma}^2,$$

then we can rewrite (B.13) as:

$$\mathrm{var}\big(\hat{\phi}_{v,\,[5]}(n)\big) = \phi_v^2(n)\frac{1}{L}\frac{1}{M-1}\left(1+\frac{\tilde{\gamma}^2}{\bar{\gamma}^2}\right). \tag{B.15}$$

Comparing (B.15) and (B.11) we can now deduce, that the MSE of $\hat{\phi}_{v,\,[5]}(n)$ can be either greater or equal to the MSE of $\hat{\phi}_{v,\,[4]}(n)$ (and the CRLB), but

can never be lower. The MSEs of these two estimators are equal only when the eigenvalues of $\tilde{\mathbf{\Gamma}}_{\mathbf{v}}$ are all equal (i.e. when $\tilde{\gamma}^2 = 0$). Since $\tilde{\mathbf{\Gamma}}_{\mathbf{v}}$ is Hermitian, it follows that for this special case to occur, $\tilde{\mathbf{\Gamma}}_{\mathbf{v}}$ must be a scaled identity matrix [13]. In all other cases, the reverberation PSD estimator from [4] outperforms the one from [5]. An important observation is that for $M = 2$ the matrix $\tilde{\mathbf{\Gamma}}_{\mathbf{v}}$ reduces to a scalar, such that $\tilde{\gamma}^2$ is always equal to zero. It follows that for $M = 2$ the reverberation PSD estimators from [4] and [5] achieve the same MSE under all possible conditions.

We can also compute the upper bound of the variance of the reverberation PSD estimator from [5]. The ratio $\tilde{\gamma}^2/\bar{\gamma}^2$ in (B.15) is maximal when all but one eigenvalue tend to zero (all energy is concentrated in a single eigenvalue). This may occur when the interference is dominated by one directional component. For such interferences the variance (and MSE) of $\hat{\phi}_{v,\,[5]}(n)$ equals:

$$\max_{\tilde{\mathbf{\Gamma}}_{\mathbf{v}}} \operatorname{var}\big(\hat{\phi}_{v,\,[5]}(n)\big) = \phi_v^2(n)\,\frac{1}{L}, \tag{B.16}$$

i.e. is $M-1$ times larger than that of $\hat{\phi}_{v,\,[4]}(n)$.

6 Experimental evaluation

We first confirm our theoretical results using a series of numerical simulations (Sec. 6.1). Additionally, we evaluate the MWF algorithms from [4] and [5] in a speech dereverberation experiment (Sec. 6.2). In both experiments the microphone array is composed of a pair of Oticon Epoq behind-the-ear hearing aids [14], each containing two microphones (i.e. $M = 4$). We have measured the RTF vectors $\mathbf{d}$ and the matrices $\mathbf{\Gamma}_{\mathbf{v}}$ in an anechoic chamber with the hearing-aids placed on a Head And Torso acoustic Simulator (HATS). The reference position for calculating $\mathbf{d}$ and $\mathbf{\Gamma}_{\mathbf{v}}$ was chosen as one of the microphones ($m = 1$), such that the corresponding elements of $\mathbf{d}$ and $\mathbf{\Gamma}_{\mathbf{v}}$ were equal to one. We used the RTF vector measured for the source position directly in front of the HATS.

6.1 Experiment 1: MSE of PSD estimation

In order to verify the theoretical results of Sec. 5, we have conducted a number of iterations of a numerical simulation. In each iteration a test signal $\mathbf{y}(n)$ was generated using $N = 25000$ pseudo-random STFT vectors drawn from a circularly-symmetric multivariate complex Gaussian distribution. The covariance matrix of this distribution was modeled according to (B.2), using the measured RTF vector $\mathbf{d}$ and matrix $\mathbf{\Gamma}_{\mathbf{v}}$ for the STFT frequency bin corresponding to 1 kHz. In each iteration $\phi_s(n)$ and $\phi_v(n)$ were set to correspond to different input SNRs between -15 and 20 dB at the reference microphone.

Next, the PSD estimators from [4] and [5] were used to estimate $\phi_s(n)$ and $\phi_v(n)$ of the test signals. The averaging length in (B.5) was set to $L = 10$ frames. Because the true values of the PSDs were known, it was possible to

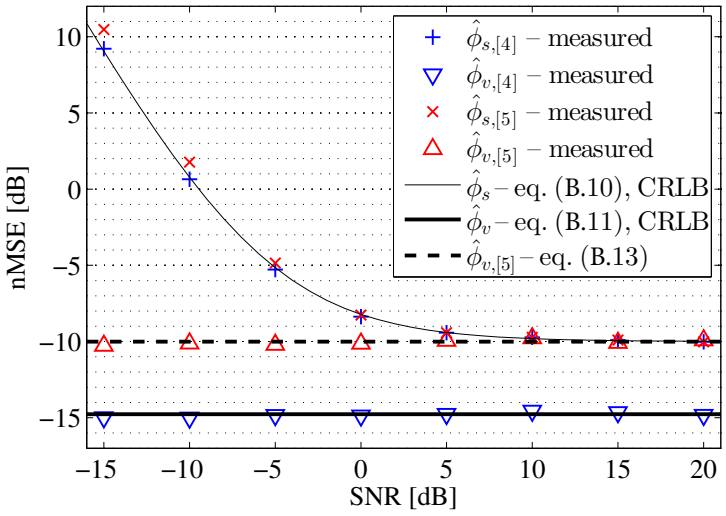


Fig. B.1: Normalized MSE of the reverberation and direct-path speech PSD estimators from [4] and [5], as a function of the input SNR, measured numerically and compared to the theoretical values. ($M = 4$, $f = 1\,\text{kHz}$, $L = 10$)

compute the MSE achieved by each of the estimators under each of the simulated SNRs. To facilitate the comparison of the obtained results, we normalized the measured MSEs by the square of the parameter of interest:

$$\text{nMSE}(\hat{\phi}_{p,r}) = \frac{\text{MSE}(\hat{\phi}_{p,r})}{\phi_p^2} = \frac{1}{N-L+1} \sum_{n=L}^{N} \frac{(\hat{\phi}_{p,r}(n) - \phi_p)^2}{\phi_p^2},$$

with p and r defined as in (B.9).

The results of this experiment are presented in Fig. B.1. For comparison, the analytically derived nMSEs formulated in (B.10), (B.11), and (B.13) are also included in the plot. The results of the numerical simulation closely agree with the theoretical formulas. The MSE achieved by the direct-path speech PSD estimator from [5] is close to, but greater than the MSE achieved by the estimator from [4]. It can also be observed that in the particular example of the simulated binaural hearing aid configuration of the microphone array, the advantage of using (B.4a) over (B.8a) for estimating the reverberation PSD is approximately 5 dB MSE for all input SNRs. Moreover, the nMSE achieved by the reverberation PSD estimator from [5] is close to the upper bound derived in (B.16), which for $L = 10$ equals $-10\,\text{dB}$ nMSE. This indicates, that the reverberation PSD estimator from [5] is not optimally suited for the simulated acoustic scenario.

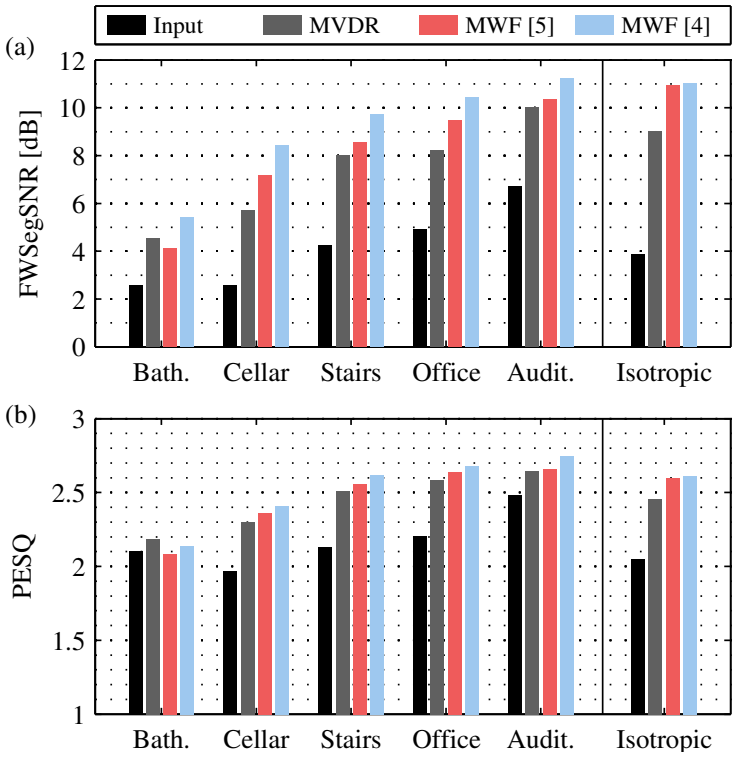


Fig. B.2: (a) FWSegSNR and (b) PESQ scores of the algorithms from [4] and [5] (denoted "MWF"). The scores computed from the unprocessed signal $y_1(n)$ ("Input"), and the output of the MVDR beamformer $\mathbf{w}_{\text{mvdr}}^H \mathbf{y}(n)$ ("MVDR") are also included.

6.2 Experiment 2: speech dereverberation performance

In order to evaluate the influence of the different PSD estimators on the MWF performance, we conducted a second simulation experiment analogous to the one presented in [4]. In this experiment the test signals were synthesized by convolving TIMIT speech sentences [15] with six different multi-channel impulse responses. Five of them were measured in real rooms using a similar microphone array as for measuring $\mathbf{d}$ and $\mathbf{\Gamma}_{\mathbf{v}}$. The sixth multi-channel impulse response (denoted "Isotropic") was synthesized to simulate an ideal cylindrically isotropic reverberant sound field. In the present study, the same room impulse responses and the same values of the non-critical simulation parameters have been used as in [4], where their detailed description may be found.

The algorithms from [4] and [5] were used to dereverberate the test signals and their performance was evaluated using the Frequency-Weighted Segmental SNR (FWSegSNR) [6] and Perceptual Evaluation of Speech Quality (PESQ) [7] objective measures. The results of this evaluation are presented in Fig. B.2. In can be observed, that the lower MSE of the PSD estimators used in the

algorithm from [4] results in a better speech dereverberation performance as measured using FWSegSNR and PESQ. Although the difference is small in the "Isotropic" condition, the advantage of using [4] over [5] increases in all realistic reverberation conditions simulated in this experiment. This suggests, that the speech dereverberation algorithm proposed in [4] may be more robust to deviations from the assumed cylindrical isotropy of the reverberation, which necessarily occur in real rooms.

7 Conclusion

In this paper we have compared two similar speech dereverberation algorithms proposed in [4] and [5]. Theoretical analysis of the direct-path speech and reverberation PSD estimators used in both algorithms revealed that for microphone numbers greater than two, the estimators used in [4] perform better than the ones used in [5] in almost all conditions. These theoretical results were confirmed in a numerical simulation.

The speech dereverberation performance of the algorithms from [4] and [5] in a four microphone binaural hearing aid configuration was measured in realistic reverberation conditions. It is found that the dereverberation algorithm from [4] outperforms [5] in terms of the FWSegSNR and PESQ objective performance measures.

References

[1] A. K. Nabelek and J. Pickett, "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech Hearing Res.*, vol. 17, no. 4, pp. 724–739, 1974.

[2] S. Doclo *et al.*, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. R. Liu, Eds. Wiley, 2008, pp. 269–302.

[3] ——, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7-8, pp. 636–656, Jul.–Aug. 2007.

[4] A. Kuklasiński *et al.*, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Proc. 22nd Eur. Signal Process. Conf. (EUSIPCO)*, Lisbon, Portugal, 2014, pp. 61–65, **(Paper A in this thesis)**.

[5] S. Braun and E. A. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. 21st Eur. Signal Process. Conf. (EUSIPCO)*, Marrakech, Morocco, 2013, pp. 1–5.

[6] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan 2008.

[7] "Perceptual evaluation of speech quality: an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Rec. P. 862*, 2001.

[8] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug 2001.

[9] H. Ye and R. D. DeGroat, "Maximum likelihood DOA estimation and asymptotic Cramér-Rao bounds for additive unknown colored noise," *IEEE Trans. Signal Process.*, vol. 43, no. 4, pp. 938–949, 1995.

[10] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Bucharest, Romania, 2012, pp. 295–299.

[11] J. Jensen and M. S. Pedersen, "Analysis of beamformer-directed single-channel noise reduction system for hearing aid applications," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brisbane, Australia, 2015, pp. 5728–5732.

[12] S. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, ser. Prentice Hall Signal Processing Series. Prentice-Hall PTR, 1993.

[13] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[14] Oticon A/S. *Oticon Epoq product brochure*. [Online]. Available: http://www.oticon.com.au/support/hearing-aids/downloads/legacy-products/~asset/cache.ashx?id=2727&type=14

[15] J. S. Garofolo *et al.*, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*. NIST, 1993.

Paper C

Maximum likelihood PSD estimation for speech enhancement in reverberation and noise

A. Kuklasiński, S. Doclo, and J. Jensen

The paper is scheduled for publication in
IEEE/ACM Transactions on Audio, Speech and Language Processing
vol. 24, no. 9, pp. 1595–1608, 2016.

Abstract

In this contribution we focus on the problem of power spectral density (PSD) estimation from multiple microphone signals in reverberant and noisy environments. The PSD estimation method proposed in this paper is based on the maximum likelihood (ML) methodology. In particular, we derive a novel ML PSD estimation scheme that is suitable for sound scenes which besides speech and reverberation consist of an additional noise component whose second-order statistics are known. The proposed algorithm is shown to outperform an existing similar algorithm in terms of PSD estimation accuracy. Moreover, it is shown numerically that the mean squared estimation error achieved by the proposed method is near the limit set by the corresponding Cramér-Rao lower bound. The speech dereverberation performance of a multi-channel Wiener filter (MWF) based on the proposed PSD estimators is measured using several instrumental measures and is shown to be higher than when the competing estimator is used. Moreover, we perform a speech intelligibility test where we demonstrate that both the proposed and the competing PSD estimators lead to similar intelligibility improvements.

1 Introduction

Reverberation and additive noise can lower the perceived quality and hinder the intelligibility of speech. This is particularly a problem in speech communication scenarios where the microphones of the receiving/recording device are at a distance from the speaker, e.g. as in hands-free telephony or in hearing aids. Clearly, noise and reverberation reduction algorithms are of practical interest.

In the literature many types of processing algorithms have been proposed for dereverberation and/or noise reduction in speech signals. Because in most scenarios both noise and reverberation are present, we focus on algorithms that can be used to jointly reduce these two types of interference (as opposed to only one of them). Moreover, we specifically focus on reduction of the late reverberation because it is believed to be particularly detrimental for speech intelligibility [1]. Following [2], speech dereverberation algorithms can be broadly divided into spectral enhancement, spatial processing, and system identification/inversion algorithms. The latter class of algorithms is generally more appropriate for dereverberation than for noise reduction (with some exceptions, e.g.: [3, 4]) and is generally used for equalization of the deterministic part of the impulse responses, rather than their stochastic (i.e. predominately late) part. On the other hand, the first two classes of algorithms (spectral enhancement and spatial processing) are well suited for noise reduction [5] and for late reverberation reduction [2]. Hence, we focus on these two types of algorithms.

Most spectral enhancement algorithms are implemented in the spectro-temporal domain and are usually based on an *a priori* statistical model of the signal components (for overviews see [5–7]). For example, in many noise re-

duction algorithms the noise power is estimated only in some spectro-temporal regions (e.g. when the signal is dominated by the noise) and is assumed to be approximately stationary between them. Speech dereverberation algorithms are mostly targeted at suppression of the late reverberation, which is often modeled as exponentially decaying and additive (e.g. [8, 9]). These and similar statistical models are used to estimate the signal-to-interference ratio in individual spectro-temporal regions, which are processed accordingly using e.g. the spectral subtraction rule or the Wiener filter [8, 9].

Spatial processing algorithms, or beamformers, work by combining the signals of an array of microphones such that it is sensitive to sounds impinging from a specific direction while suppressing sounds from other directions. Obviously, beamformers are only effective in scenarios where the interference (noise and/or reverberation) impinges on the microphone array from different directions than the target speech.

Spectral enhancement and beamforming algorithms are often combined to create a two-step algorithm where the beamformer is followed by a single channel spectral enhancement scheme (in this context referred to as the post-filter). Among the first methods of this type proposed for speech dereverberation and noise reduction were [10, 11], both composed of a delay-and-sum beamformer and a coherence-based post-filter. The beamformers and post-filters in algorithms proposed more recently are generally based on some optimality criteria, most notably the linear minimum mean square error (MMSE) resulting in the multi-channel Wiener filter (MWF) [12, 13]. The MWF depends on the inter-microphone covariance matrices of the desired (target speech) and of the interference (noise and late reverberation) components of the input signal. These matrices are usually not known but in some scenarios their structure can be modeled such that only few parameters remain to be estimated. In this paper we employ a set of assumptions that result in a signal covariance model where only the power spectral densities (PSDs) of the target speech and of the late reverberation need to be estimated.

Several methods exist for estimating the speech and the late reverberation PSDs in the considered setup. Estimators operating on a single microphone signal are generally considered inferior to PSD estimators using multiple microphones [14]. In the past, multi-microphone estimators based on the inter-microphone coherence have been proposed [10, 11]. These estimators generally are based on the assumption that the late reverberation is uncorrelated between microphones (invalid e.g. for low frequencies and finite inter-microphone distance). More recently, estimators based on optimality criteria have been proposed, e.g. by Braun and Habets [15], and by the authors of this study [16]. Both these estimators are based on the maximum likelihood (ML) methodology, and have been compared with respect to the estimation accuracy in [17]. For the special case where the signals are composed of only speech and reverberation, the estimator from [16] has been found to yield superior statistical performance compared with the estimator from [15]. In fact, in [14] it was argued that the estimator used in [16] is optimal in the minimum variance

unbiased (MVU) sense.

A disadvantage of the estimator in [16] compared to the estimator in [15] is that the former does not take the additive noise into account. On the other hand, the estimator in [15] is derived using an unrealistic statistical assumption which results in its decreased estimation performance [17]. In this contribution we propose a scheme which avoids both these limitations. Specifically, we propose a novel multi-microphone PSD estimator which is approximately ML-optimal and generalizes the method from [16] to signal models including a target signal contaminated by late reverberation and additive noise.

This paper is structured as follows. Section 2 presents the signal model and discusses the employed statistical assumptions. In Section 3 the proposed estimator is derived and several practically relevant special cases are presented for which the estimator is particularly simple. In Section 4 a detailed experimental evaluation is performed and the statistical performance of the proposed estimator is compared to the estimator from [15]. It is also shown, that the mean squared error of the proposed PSD estimator is close to the lowest possible for unbiased estimators, as set by the Cramér-Rao lower bound (CRLB). In Section 5 the speech dereverberation performance of an MWF based on the two compared PSD estimation methods is evaluated in terms of: the frequency-weighted segmental signal-to-noise ratio (FWSegSNR) [18], the perceptual evaluation of speech quality (PESQ) [19] measure, two interference attenuation, and one speech distortion measure [20, 21]. Lastly, in Section 6, the two variants of the MWF are evaluated in a speech intelligibility (SI) test with human subjects. Section 7 concludes the paper.

2 Signal model and statistical assumptions

Consider an array of M microphones in a reverberant room where a single talker is active. Speech generated by the talker reaches the microphones not only via the direct propagation path, but also via multiple reflections off the walls and other surfaces in the room. In most practical situations the microphone signals are further disrupted by the microphone self-noise and by other additive noise sources.

For a particular arrangement of a sound source and a sound receiver, acoustic properties of a room can be compactly expressed in terms of a room impulse response (RIR). We adopt an often-made assumption that RIRs are composed of three distinct parts: the direct path response, the early reflections, and the late reverberation. The direct and early components of reverberant speech are generally considered advantageous for speech intelligibility [1]; hence, we refer to their sum as the target signal. In specific scenarios it might not be desirable or practical to include all early reflections (conventionally the first 50 ms of the RIR) in the target signal model. For this reason we define the target signal as the direct path speech plus those of its early reflections whose delay relative to the direct path is less than a certain threshold t_s. The remaining

early reflections are not accounted for in the signal model. All other components of the signal, i.e. the late reverberation, the microphone self-noise, and other additive noise types, are all considered an interference because of their detrimental effect on speech quality and intelligibility.

Let $y_m(t)$ denote the time-domain signal of the m-th microphone of the array ($m = 1, \ldots, M$), where t is a discrete time index. Due to the wide-band and non-stationary nature of the speech, it is often convenient to implement speech processing algorithms in the spectro-temporal domain. Thus, we express $y_m(t)$ as its short time Fourier transform (STFT) given by:

$$y_m(k,n) = \sum_{t=0}^{T-1} y_m(t+nD)w(t)e^{-2\pi i k \frac{t}{T}},$$

where k is the frequency bin index, n is the time frame index, the STFT length is denoted by T, the filterbank decimation factor is denoted by D, and $w(t)$ is the analysis window function. For notational conciseness we stack the STFT coefficients corresponding to all of the microphones in a vector $\mathbf{y}(k,n) = [\, y_1(k,n) \ldots \, y_M(k,n) \,]^T$. Furthermore, we assume that $\mathbf{y}(k,n)$ is a sum of three components:

$$\mathbf{y}(k,n) = \mathbf{s}(k,n) + \mathbf{r}(k,n) + \mathbf{x}(k,n), \tag{C.1}$$

where $\mathbf{s}(k,n)$ corresponds to the target signal, $\mathbf{r}(k,n)$ corresponds to the late reverberation, and $\mathbf{x}(k,n)$ is the additive noise component (i.e. sum of the microphone self-noise, ambient noise, and possibly other additive interferences).

We assume that $\mathbf{y}(k,n)$ is uncorrelated across frequency bins, which allows us to omit the frequency bin index k in the subsequent presentation. All processing is performed independently in all frequency bins. Moreover, for mathematical tractability, we assume that $\mathbf{y}(n)$ is uncorrelated across time frames. In other words, we neglect the influence of any existing overlap between the time frames and any autocorrelation the microphone signals may exhibit for delays larger than the STFT length. Because reverberant speech signals are autocorrelated and the time frames do overlap, this assumption is, at best, only approximately valid. Nevertheless, it is employed in many speech processing algorithms (e.g. [8, 15, 22]) and the general success of these methods reflects that it is a useful working assumption.

Because the additive noise is generated by physical processes independent of the speech, we assume that $\mathbf{x}(n)$ is uncorrelated with $\mathbf{s}(n)$ and $\mathbf{r}(n)$. Moreover, we assume that the late reverberation $\mathbf{r}(n)$ is uncorrelated with the target signal $\mathbf{s}(n)$. This is an often used assumption (e.g. [8, 9, 15]), which can be justified by the fact that the late part of RIRs is disturbed by thermal fluctuations of the air [23] and slight movements of the source and the microphone array [24] which are unavoidable in practical scenarios. Moreover, in applications where the STFT length has to be very short (such as in hearing aids), in any time frame the reverberation can be argued to be correlated mostly with the speech component of the preceding time frames, but not of the current one.

The covariance matrix of $\mathbf{y}(n)$ is defined as:

$$\mathbf{\Phi}_{\mathbf{y}}(n) = E[\mathbf{y}(n)\mathbf{y}^H(n)], \tag{C.2}$$

where $E[\cdot]$ denotes the expectation operator and $(\cdot)^H$ is the Hermitian transpose. Each of the diagonal elements of $\mathbf{\Phi}_{\mathbf{y}}(n)$ is equal (up to a normalization constant) to the power spectral density (PSD) of the respective microphone signal in the particular frequency bin. Similarly, off-diagonal elements of $\mathbf{\Phi}_{\mathbf{y}}(n)$ correspond to the cross-PSDs between the respective microphones. Hence, we refer to $\mathbf{\Phi}_{\mathbf{y}}(n)$ as the cross-PSD matrix of $\mathbf{y}(n)$. Because we assume that the signal components are uncorrelated, $\mathbf{\Phi}_{\mathbf{y}}(n)$ can be decomposed into a sum of cross-PSD matrices of the individual signal components. Hence:

$$\mathbf{\Phi}_{\mathbf{y}}(n) = \mathbf{\Phi}_{\mathbf{s}}(n) + \mathbf{\Phi}_{\mathbf{r}}(n) + \mathbf{\Phi}_{\mathbf{x}}(n), \tag{C.3}$$

where $\mathbf{\Phi}_{\mathbf{s}}(n)$, $\mathbf{\Phi}_{\mathbf{r}}(n)$, and $\mathbf{\Phi}_{\mathbf{x}}(n)$ denote the cross-PSD matrices of $\mathbf{s}(n)$, $\mathbf{r}(n)$, and $\mathbf{x}(n)$, respectively.

We assume that the STFT coefficients of the microphone signal and its individual components are circularly-symmetric complex Gaussian distributed, e.g: $\mathbf{y}(n) \sim \mathcal{N}_C(\mathbf{0}, \mathbf{\Phi}_{\mathbf{y}}(n))$. While it is known that the STFT coefficients, particularly of the speech component, are more accurately modeled using super-Gaussian distributions (see e.g. [25–27]), the resulting estimators tend to become significantly more complicated (see e.g. [21]). Thus, the Gaussian assumption appears to be a good tradeoff between accuracy and mathematical tractability.

We model the talker as a single point-source. The direct path and the early reflections can be modeled as linear filters acting on the speech emitted by the talker. In effect, the target signal received by any of the microphones is a linearly filtered version of the target signal anywhere else in the room. In order to use this property, we select a certain reference position (conventionally one of the microphones) and denote the STFT of the target signal at that position by $s(n)$ (a scalar). Next, we let $\mathbf{d}$ denote a vector of relative transfer functions (RTFs) [28] of the target signal from the chosen reference position to all of the microphones (evaluated at the center frequency of the current frequency bin). For $\mathbf{d}$ to represent the RTFs accurately, the early reflection threshold t_s must be shorter than the STFT length. Using the above definitions, we can write:

$$\mathbf{s}(n) = s(n)\mathbf{d}. \tag{C.4}$$

We assume that an estimate of $\mathbf{d}$ is available (e.g. because the application at hand allows its accurate off-line estimation, or, alternatively, by use of an on-line estimation scheme such as [29, 30]). Using (C.4) in the definition of $\mathbf{\Phi}_{\mathbf{s}}(n)$ results in:

$$\mathbf{\Phi}_{\mathbf{s}}(n) = E[\mathbf{s}(n)\mathbf{s}^H(n)] = \phi_s(n)\mathbf{d}\mathbf{d}^H. \tag{C.5}$$

It follows that the matrix $\mathbf{\Phi}_{\mathbf{s}}(n)$ is rank-one and constant up to a scaling factor $\phi_s(n)$, which denotes the time-varying PSD of the target speech at the reference position.

The late reverberation cross-PSD matrix may be written as:

$$\mathbf{\Phi_r}(n) = \phi_r(n)\mathbf{\Gamma_r}, \tag{C.6}$$

where $\phi_r(n)$ denotes the time-varying (scalar) PSD of the late reverberation at the reference position and $\mathbf{\Gamma_r}$ is the cross-PSD matrix of the late reverberation normalized by $\phi_r(n)$. The proposed method is based on the assumption that $\mathbf{\Gamma_r}$ is full-rank and known, or, equivalently, that the spatial distribution of the late reverberation is known. Drawing from statistical models employed in theoretical room acoustics (see e.g. [31]) we assume that all directions contribute equally to the late reverberant sound field, i.e. that this sound field is isotropic. In consequence, $\mathbf{\Gamma_r}$ can be measured *a priori* as it does not depend on the position or orientation of the microphone array within the room. For free-field microphone arrays, $\mathbf{\Gamma_r}$ can even be calculated analytically using information on the microphone array geometry [32, 33]. For other microphone arrays, $\mathbf{\Gamma_r}$ has to be measured or modeled numerically. In many rooms, the floor and the ceiling are the most acoustically damped surfaces. In effect, the vertical component of the reverberant sound field is damped more than its horizontal components. In such rooms the reverberation is more accurately modeled as cylindrically, rather than spherically isotropic.

We assume that the third component of the signal model, $\mathbf{x}(n)$, is related to an additive noise whose statistics are varying slowly—a realistic assumption if $\mathbf{x}(n)$ is used to model the sum of the noise generated by the microphones and by other sources: ambiance, ventilation equipment, car or airplane cabin noise, etc. As a consequence, the cross-PSD matrix $\mathbf{\Phi_x}$ can be assumed approximately constant across short spans of time (hence, we omit index n). We assume that $\mathbf{\Phi_x}$ is known or that a reliable estimate thereof is available. In practice, an estimation scheme such as the multi-microphone speech probability estimator proposed in [34] could be used to periodically update $\mathbf{\Phi_x}$ during time-frequency regions where speech and late reverberation levels are low compared to that of the noise (e.g. between speech utterances).

Using (C.5) and (C.6), the overall model for the microphone input cross-PSD matrix can be re-written as (cf. (C.3)):

$$\mathbf{\Phi_y}(n) = \phi_s(n)\mathbf{d}\mathbf{d}^H + \phi_r(n)\mathbf{\Gamma_r} + \mathbf{\Phi_x}. \tag{C.7}$$

In this model only the scalar PSDs $\phi_s(n)$ and $\phi_r(n)$ are unknown; their estimation and application to speech dereverberation is the focus of this paper. To facilitate the derivation of the proposed estimators, we assume that $\phi_s(n)$ and $\phi_r(n)$ can be considered approximately constant across a certain number L of consecutive time frames of the STFT. For small L, such that L frames span less than 50 ms, this is analogous to the commonly made assumption of short-time speech stationarity.

The proposed PSD estimation method is intended for reverberant and noisy speech signals, and the employed assumptions are motivated by this application. However, the proposed algorithm is equally useful for other types of

signals, provided that the assumptions made are satisfied, i.e. that the signals are approximately Gaussian and that their cross-PSD matrix can be modeled using (C.7).

3 Derivation of the proposed PSD estimators

In this section we derive the proposed maximum likelihood estimators (MLEs) of $\phi_s(n)$ and $\phi_r(n)$. We begin by formulating a probability density function (PDF) of the input signal $\mathbf{y}(n)$, which we subsequently use to define a joint likelihood function of $\phi_s(n)$ and $\phi_r(n)$.

Due to the assumptions outlined in Section 2, the input signal vectors $\mathbf{y}(n)$ in any L consecutive time frames can be considered approximately independent and identically distributed. It follows, that the joint PDF of the signal in these L time frames can be calculated as the product of the PDFs of $\mathbf{y}(n)$ in individual time frames. Denoting the sample cross-PSD matrix of the input signal as:

$$\hat{\mathbf{\Phi}}_{\mathbf{y}}(n) = \frac{1}{L}\sum_{l=0}^{L-1}\mathbf{y}(n-l)\mathbf{y}^H(n-l), \tag{C.8}$$

we can compactly express the joint complex Gaussian PDF of $\mathbf{y}(n)$ in L consecutive time frames as:

$$f = \frac{1}{\pi^{LM}|\mathbf{\Phi}_{\mathbf{y}}(n)|^L}\exp\big[-L\,\mathrm{tr}\big(\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)\mathbf{\Phi}_{\mathbf{y}}^{-1}(n)\big)\big]. \tag{C.9}$$

This joint PDF depends on ϕ_s and ϕ_r (through $\mathbf{\Phi}_{\mathbf{y}}(n)$, cf. (C.7)), which are regarded as deterministic but unknown.

The required joint likelihood function is obtained by interpreting the joint PDF (C.9) as a function of ϕ_s and ϕ_r. For mathematical convenience we will be operating on its natural logarithm $\mathcal{L} = \log(f)$. Omitting one non-essential term $(-LM\log(\pi))$, this log-likelihood $\mathcal{L}$ can be written as:

$$\mathcal{L}(\phi_s,\phi_r) = -L\log|\mathbf{\Phi}_{\mathbf{y}}(n)| - L\,\mathrm{tr}\big[\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)\mathbf{\Phi}_{\mathbf{y}}^{-1}(n)\big], \tag{C.10}$$

where tr[·] denotes the matrix trace operator. The MLEs of $\phi_s(n)$ and $\phi_r(n)$ are defined as the coordinates of the global maximum of $\mathcal{L}(\phi_s,\phi_r)$ and can be found by solving a two-dimensional optimization problem:

$$\big(\hat{\phi}_{s,\mathrm{ML}}(n),\,\hat{\phi}_{r,\mathrm{ML}}(n)\big) = \arg\max_{\phi_s,\,\phi_r}\mathcal{L}(\phi_s,\phi_r), \tag{C.11}$$

where $\hat{\phi}_{s,\mathrm{ML}}(n)$ and $\hat{\phi}_{r,\mathrm{ML}}(n)$ denote the MLEs of $\phi_s(n)$ and $\phi_r(n)$, respectively.

3.1 Estimator of the target speech PSD

As shown in [35], the MLE of $\phi_s(n)$ can be analytically found by maximizing the likelihood function (C.10) conditioned on $\hat{\phi}_{r,\mathrm{ML}}(n)$, i.e. by solving a one-dimensional optimization problem (cf. (C.11)):

$$\hat{\phi}_{s,\mathrm{ML}}(n) = \arg\max_{\phi_s} \mathcal{L}(\phi_s; \hat{\phi}_{r,\mathrm{ML}}).$$

Let $\hat{\mathbf{\Phi}}_{\mathbf{v}}(n) = \hat{\phi}_{r,\mathrm{ML}}(n)\mathbf{\Gamma}_{\mathbf{r}} + \mathbf{\Phi}_{\mathbf{x}}$ denote the MLE of the cross-PSD matrix of the total interference. Then, the MLE $\hat{\phi}_{s,\mathrm{ML}}(n)$ can be written as [35, Appendix B]:

$$\hat{\phi}_{s,\mathrm{ML}}(n) = \mathbf{w}_{\mathrm{MVDR}}^{H}(n)\big[\hat{\mathbf{\Phi}}_{\mathbf{y}}(n) - \hat{\mathbf{\Phi}}_{\mathbf{v}}(n)\big]\mathbf{w}_{\mathrm{MVDR}}(n), \tag{C.12}$$

where

$$\mathbf{w}_{\mathrm{MVDR}}(n) = \frac{\hat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(n)\mathbf{d}}{\mathbf{d}^{H}\hat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(n)\mathbf{d}} \tag{C.13}$$

is the weight vector of a minimum variance distortionless response (MVDR) beamformer [36]. The MLE (C.12) is a function of (is conditioned on) $\hat{\phi}_{r,\mathrm{ML}}(n)$ and can be interpreted as the difference between the estimates of the total PSD and the interference PSD at the output of the MVDR beamformer.

3.2 Estimator of the late reverberation PSD

Because $\hat{\phi}_{s,\mathrm{ML}}(n)$ and $\hat{\phi}_{r,\mathrm{ML}}(n)$ are analytically related by (C.12), a one-dimensional, concentrated likelihood function of ϕ_r can be defined as: $\mathcal{L}'(\phi_r) = \mathcal{L}\big(\hat{\phi}_{s,\mathrm{ML}}(\phi_r), \phi_r\big)$. The exact MLE of $\phi_r(n)$ can be found as the argument of the maximum of $\mathcal{L}'(\phi_r)$ [35]. Unfortunately, for the signal model at hand this optimization problem is not easily tractable. Instead of resorting to numerical optimization methods to find the maximum of $\mathcal{L}'(\phi_r)$, we propose a simplified MLE of $\phi_r(n)$ using a modified form of the input signal model.

The modifications consist of two steps. First, we pass the input STFT vector $\mathbf{y}(n)$ through a target-blocking matrix $\mathbf{B} \in \mathbb{C}_{M\times(M-1)}$ defined as [37]:

$$[\mathbf{B}\ \mathbf{b}] = \mathbf{I} - \mathbf{d}(\mathbf{d}^{H}\mathbf{d})^{-1}\mathbf{d}^{H}, \tag{C.14}$$

where $\mathbf{B}$ denotes the first $M-1$ columns and $\mathbf{b}$ denotes the last column of the matrix on the right-hand-side of (C.14). The columns of $\mathbf{B}$ can be interpreted as a set of $M-1$ target-canceling beamformers, i.e.: $\mathbf{B}^{H}\mathbf{s}(n) = \mathbf{0}$. Hence, the blocked input signal can be written as: $\mathbf{B}^{H}\mathbf{y}(n) = \mathbf{B}^{H}\mathbf{r}(n) + \mathbf{B}^{H}\mathbf{x}(n)$ (cf. (C.1)), and its cross-PSD matrix as (cf. (C.2)):

$$\begin{aligned} E[\mathbf{B}^{H}\mathbf{y}(n)\mathbf{y}^{H}(n)\mathbf{B}] &= \mathbf{B}^{H}\mathbf{\Phi}_{\mathbf{y}}(n)\mathbf{B} \\ &= \mathbf{B}^{H}\mathbf{\Phi}_{\mathbf{r}}(n)\mathbf{B} + \mathbf{B}^{H}\mathbf{\Phi}_{\mathbf{x}}\mathbf{B}. \end{aligned}$$

The second modification of the signal model has the objective of diagonalizing $\mathbf{B}^H\mathbf{\Phi_x}\mathbf{B}$, i.e. the additive noise component of the blocked input cross-PSD matrix. To that end, we use a whitening matrix $\mathbf{D} \in \mathbb{C}_{(M-1)\times(M-1)}$ and define it as the Cholesky factor of the inverse of $\mathbf{B}^H\mathbf{\Phi_x}\mathbf{B}$:

$$\mathbf{D}\mathbf{D}^H = (\mathbf{B}^H\mathbf{\Phi_x}\mathbf{B})^{-1}. \tag{C.15}$$

It is necessary to assume that $\mathbf{B}^H\mathbf{\Phi_x}\mathbf{B}$ is full rank. N.b.: it is sufficient that real (and, therefore, noisy) microphones are used in the array to guarantee that $\mathbf{B}^H\mathbf{\Phi_x}\mathbf{B}$ is full rank, even if the other noise types contributing to $\mathbf{x}(n)$ (e.g. ambient noise) do not by themselves result in a full rank cross-PSD matrix.

The blocked and whitened signal is given by $\tilde{\mathbf{y}}(n) = \mathbf{D}^H\mathbf{B}^H\mathbf{y}(n)$ and its cross-PSD matrix can be found as:

$$\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n) = \mathbf{D}^H\mathbf{B}^H\mathbf{\Phi_y}(n)\mathbf{B}\mathbf{D} = \phi_r(n)\mathbf{\Gamma}_{\tilde{\mathbf{r}}} + \mathbf{I}, \tag{C.16}$$

where $\mathbf{\Gamma}_{\tilde{\mathbf{r}}} = \mathbf{D}^H\mathbf{B}^H\mathbf{\Gamma_r}\mathbf{B}\mathbf{D}$.

As a result of the described modifications, the matrix $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ exhibits a useful feature: its eigenvectors are the same as that of the matrix $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$. Equivalently, the eigendecompositions of $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ and $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$ use the same unitary matrix $\mathbf{U}$:

$$\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n) = \mathbf{U}\mathbf{\Lambda_\Phi}(n)\mathbf{U}^H, \quad \mathbf{\Gamma}_{\tilde{\mathbf{r}}} = \mathbf{U}\mathbf{\Lambda_\Gamma}\mathbf{U}^H, \tag{C.17}$$

where the orthonormal columns of $\mathbf{U}$ are the eigenvectors, and where $\mathbf{\Lambda_\Phi}(n)$ and $\mathbf{\Lambda_\Gamma}$ are diagonal matrices of the eigenvalues of $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ and $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$, respectively. Because $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$ is constant, so are $\mathbf{U}$ and $\mathbf{\Lambda_\Gamma}$. Due to (C.16), $\mathbf{\Lambda_\Phi}(n)$ and $\mathbf{\Lambda_\Gamma}$ are related as:

$$\mathbf{\Lambda_\Phi}(n) = \phi_r(n)\mathbf{\Lambda_\Gamma} + \mathbf{I}. \tag{C.18}$$

Equivalently: $\lambda_{\mathbf{\Phi},m} = \phi_r(n)\lambda_{\mathbf{\Gamma},m} + 1$, where $\lambda_{\mathbf{\Phi},m}$ and $\lambda_{\mathbf{\Gamma},m}$ denote the m-th eigenvalue of $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ and $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$, respectively.

Using the blocked and whitened signal model (C.16) we can formulate a new and simplified log-likelihood of ϕ_r. It has a form analogous to (C.10) with the input cross-PSD matrix and its estimate substituted by their blocked and whitened counterparts $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ and $\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)$:

$$\mathcal{L}^{''}(\phi_r) = -L\log|\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)| - L\,\mathrm{tr}\big[\mathbf{\Phi}_{\tilde{\mathbf{y}}}^{-1}(n)\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\big]. \tag{C.19}$$

The proposed MLE of ϕ_r is defined as: $\hat{\phi}_r(n) = \arg\max_{\phi_r}\mathcal{L}^{''}(\phi_r)$. To find $\hat{\phi}_r(n)$ we must first find the derivative of $\mathcal{L}^{''}(\phi_r)$ with respect to ϕ_r. We compute it by using the fact that for any invertible matrix $\mathbf{A}(\theta)$ the following identities hold ($\mathbf{A}(\theta)$ is a function of θ) [38, 39]:

$$\frac{\mathrm{d}\log|\mathbf{A}(\theta)|}{\mathrm{d}\theta} = \mathrm{tr}\left[\mathbf{A}^{-1}(\theta)\frac{\mathrm{d}\mathbf{A}(\theta)}{\mathrm{d}\theta}\right],$$

$$\frac{\mathrm{d}\,\mathrm{tr}\left[\mathbf{A}^{-1}(\theta)\mathbf{Z}\right]}{\mathrm{d}\theta} = -\,\mathrm{tr}\left[\mathbf{A}^{-1}(\theta)\frac{\mathrm{d}\mathbf{A}(\theta)}{\mathrm{d}\theta}\mathbf{A}^{-1}(\theta)\mathbf{Z}\right].$$

We also note that the derivative of $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ with respect to ϕ_r is equal to $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$ (cf. (C.16)). The (known) result is [35, Eq. (2)]:

$$\frac{\mathrm{d}\mathcal{L}^{''}(\phi_r)}{\mathrm{d}\phi_r} = -L\,\mathrm{tr}\big[\mathbf{\Phi}_{\tilde{\mathbf{y}}}^{-1}(n)\mathbf{\Gamma}_{\tilde{\mathbf{r}}} - \mathbf{\Phi}_{\tilde{\mathbf{y}}}^{-1}(n)\mathbf{\Gamma}_{\tilde{\mathbf{r}}}\mathbf{\Phi}_{\tilde{\mathbf{y}}}^{-1}(n)\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\big]. \tag{C.20}$$

The proposed estimator is found by setting (C.20) to zero and solving for ϕ_r. To do so, we re-write (C.20) using (C.17):

$$\mathrm{tr}\big[\mathbf{\Lambda}_{\mathbf{\Phi}}^{-1}(n)\mathbf{\Lambda}_{\mathbf{\Gamma}} - \mathbf{\Lambda}_{\mathbf{\Phi}}^{-1}(n)\mathbf{\Lambda}_{\mathbf{\Gamma}}\mathbf{\Lambda}_{\mathbf{\Phi}}^{-1}(n)\mathbf{U}^H\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\mathbf{U}\big] = 0.$$

Exploiting the diagonal structure of the involved matrices and using (C.18), this can be written as:

$$\sum_{m=1}^{M-1}\left[\frac{\lambda_{\mathbf{\Gamma},m}}{(\phi_r\lambda_{\mathbf{\Gamma},m}+1)} - \frac{\lambda_{\mathbf{\Gamma},m}\,g_m(n)}{(\phi_r\lambda_{\mathbf{\Gamma},m}+1)^2}\right] = 0, \tag{C.21}$$

where $g_m(n)$ denotes the m-th diagonal element of $\mathbf{U}^H\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\mathbf{U}$. It can be seen that (C.21) is a sum of $2(M-1)$ rational terms. By converting all these terms to a common denominator ($\prod_{k=1}^{M-1}(\phi_r\lambda_{\mathbf{\Gamma},m}+1)^2$), taking only the resulting numerators into account, and some additional simplifications, (C.21) can be expressed as a sum of $M-1$ polynomials in ϕ_r:

$$p(\phi_r) = \sum_{m=1}^{M-1} p_m(\phi_r), \quad \text{where} \tag{C.22}$$

$$p_m(\phi_r) = \underbrace{\left(\phi_r - \frac{g_m(n)-1}{\lambda_{\mathbf{\Gamma},m}}\right)}_{\text{order } 1}\underbrace{\prod_{k=1}^{M-1,\,k\neq m}\left(\phi_r + \frac{1}{\lambda_{\mathbf{\Gamma},k}}\right)^2}_{\text{order } 2(M-2)}.$$

The polynomial $p(\phi_r)$ is of odd order: $2M-3$. Hence, at least 1 and at most $2M-3$ real roots of $p(\phi_r)$ exist. When more than one real root of $p(\phi_r)$ exists, the one yielding the highest value of the likelihood (C.19) must be chosen as the MLE $\hat{\phi}_r(n)$.

For convenience, a pseudo-code representation of the algorithm for computing the proposed PSD estimators is provided in Figure C.1. As we show in Appendix A, usually only one real root of $p(\phi_r)$ exists. Therefore, in most cases the condition in Figure C.1, line 13 is satisfied, and it is not necessary to compute the numerical value of the likelihood (C.19).

In general, numerical methods must be applied to find the roots of $p(\phi_r)$ as no closed-form solution appears obtainable. For microphone arrays with few microphones, such as often found in hearing aids, this is computationally trivial. For large microphone arrays, solving (C.22) may become problematic in applications where computing power is limited.

Define: $\mathbf{d}$, $\mathbf{\Gamma_r}$, $\mathbf{\Phi_x}$

$$[\mathbf{B}\ \mathbf{b}] = \mathbf{I} - \mathbf{d}(\mathbf{d}^H\mathbf{d})^{-1}\mathbf{d}^H \tag{C.14}$$

$$\mathbf{D}\mathbf{D}^H = (\mathbf{B}^H\mathbf{\Phi_x}\mathbf{B})^{-1} \tag{C.15}$$

$$\mathbf{\Gamma_{\tilde{r}}} = \mathbf{D}^H\mathbf{B}^H\mathbf{\Gamma_r}\mathbf{B}\mathbf{D} \tag{C.16}$$

$$\mathbf{U}\mathbf{\Lambda_\Gamma}\mathbf{U}^H = \mathbf{\Gamma_{\tilde{r}}} \qquad \text{such that:} \qquad \mathbf{U}\mathbf{U}^H = \mathbf{I} \tag{C.17}$$

$\lambda_{\mathbf{\Gamma},m} = [\mathbf{\Lambda_\Gamma}]_{m,m}$

for all n **do**

  Define: $\mathbf{y}(n)$

  Update: $\hat{\mathbf{\Phi}}_\mathbf{y}(n)$ (C.8)

  $g_m(n) = [\mathbf{U}^H\mathbf{D}^H\mathbf{B}^H\hat{\mathbf{\Phi}}_\mathbf{y}(n)\mathbf{B}\mathbf{D}\mathbf{U}]_{m,m}$

  Define: $p(\phi_r)$ (C.22)

  $\mathcal{P}(n) = \{\phi_r : p(\phi_r) = 0\}$

  if $\big|\mathcal{P}(n)\big| = 1$ **then**

    $\hat{\phi}_r(n) = \{\mathcal{P}(n)\}$

  else

    $\hat{\phi}_r(n) = \arg\max_{\phi_r \in \mathcal{P}(n)} \mathcal{L}''(\phi_r)$ (C.19)

  end if

  $\hat{\mathbf{\Phi}}_\mathbf{v}(n) = \hat{\phi}_r(n)\mathbf{\Gamma_r} + \mathbf{\Phi_x}$

  $\mathbf{w}_{\mathrm{MVDR}}(n) = \hat{\mathbf{\Phi}}_\mathbf{v}^{-1}(n)\mathbf{d}\big[\mathbf{d}^H\hat{\mathbf{\Phi}}_\mathbf{v}^{-1}(n)\mathbf{d}\big]^{-1}$ (C.13)

  $\hat{\phi}_s(n) = \mathbf{w}_{\mathrm{MVDR}}^H(n)\big[\hat{\mathbf{\Phi}}_\mathbf{y}(n) - \hat{\mathbf{\Phi}}_\mathbf{v}(n)\big]\mathbf{w}_{\mathrm{MVDR}}(n)$ (C.12)

end for

Fig. C.1: A pseudocode representation of the proposed PSD estimation method. The presented routine is to be applied in all frequency bins (possibly in parallel). The set of roots of the polynomial $p(\phi_r)$ in the n-th time frame is denoted as $\mathcal{P}(n)$, with $|\mathcal{P}(n)|$ being its cardinality (number of elements). Relevant equation numbers are provided for cross-reference.

The proposed late reverberation PSD estimator $\hat{\phi}_r(n)$ is the exact MLE of $\phi_r(n)$ *in the blocked signal domain* (C.16) (to within the precision of the root-finding algorithm). However, numerical simulations indicated that $\hat{\phi}_r(n)$ is not equal to the MLE $\hat{\phi}_{r,\mathrm{ML}}(n)$ defined in (C.11), i.e. in the unmodified signal domain (C.7). This is due to the loss of information about the signal induced by the blocking operation. Additionally, the target speech PSD estimator computed according to (C.12) but conditioned on $\hat{\phi}_r(n)$ instead of $\hat{\phi}_{r,\mathrm{ML}}(n)$ is not equal to the exact MLE $\hat{\phi}_{s,\mathrm{ML}}(n)$. Therefore, both proposed PSD estimators are only approximations of the true MLE in the unmodified signal domain. Nevertheless, experimental results reported in Section 4 show that the loss of the estimation performance is very small.

3.3 Estimator of the late reverberation PSD for $\mathbf{x}(n) = \mathbf{0}$

A special case of the proposed late reverberation PSD estimator can be derived for signals where $\mathbf{x}(n) = \mathbf{0}$. Because $\mathbf{\Phi}_{\mathbf{x}} = \mathbf{0}$, the whitening operation is undefined and must be omitted. It follows, that (C.16) has to be re-written as $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n) = \phi_r(n)\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$. Using this in (C.20) a new equation for the MLE is found:

$$\phi_r^{-1}(n)\,\mathrm{tr}\big[\mathbf{I} - \mathbf{\Phi}_{\tilde{\mathbf{y}}}^{-1}(n)\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\big] = 0.$$

Unlike in the general scenario, in this special case a closed form solution for the MLE exists:

$$\hat{\phi}_{r|\mathbf{x}=\mathbf{0}}(n) = \frac{1}{M-1}\,\mathrm{tr}\big[\mathbf{\Gamma}_{\tilde{\mathbf{r}}}^{-1}\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\big]. \tag{C.23}$$

This expression can be recognized as the multi-microphone noise PSD estimator proposed in [37]. In [14] this estimator has been shown to be minimum variance unbiased (MVU). Furthermore, (an equivalent form of) the estimator (C.23) was used for late reverberation PSD estimation in an earlier paper [16] by the authors of this study.

Although the assumption that $\mathbf{x}(n) = \mathbf{0}$ often does not hold in practical applications, it is approximately satisfied in scenarios where the additive noise $\mathbf{x}(n)$ is negligible compared to the late reverberation $\mathbf{r}(n)$. In some applications, the benefits of using a closed-form estimator like (C.23) may outweigh the benefits of modeling the signal more accurately.

3.4 Estimator of the late reverberation PSD for $M = 2$

Another special case may be considered for devices with only two microphones, such as some hearing aids, smartphones, and laptops. Because the blocking matrix reduces the dimensionality of the signal by one, all vectors and matrices involved in the estimation of $\phi_r(n)$ degenerate into scalars. Then, the polynomial (C.22) degenerates into a linear equation which is easily solved:

$$\hat{\phi}_{r|M=2}(n) = \frac{g(n)-1}{\lambda_{\mathbf{\Gamma}}} = (\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n) - \mathbf{\Phi}_{\tilde{\mathbf{x}}})\mathbf{\Gamma}_{\tilde{\mathbf{r}}}{}^{-1}. \tag{C.24}$$

Note that this equation is composed of scalars; we maintain the bold print for the sake of notation continuity. For $M = 2$, the proposed late reverberation PSD estimator (C.24) and the one proposed by Braun and Habets in [15] are equivalent (can be written as identical equations).

4 Evaluation of the proposed PSD estimator in terms of the normalized mean squared error

In this section we evaluate the proposed PSD estimator and compare it with the estimator proposed by Braun and Habets [15]. As the performance metric we use the normalized mean-squared error (MSE) of estimation defined as:

$$\mathrm{nMSE}_{\phi_s} = \frac{E\big[(\hat{\phi}_s - \phi_s)^2\big]}{\phi_s^2}, \quad \mathrm{nMSE}_{\phi_r} = \frac{E\big[(\hat{\phi}_r - \phi_r)^2\big]}{\phi_r^2}. \tag{C.25}$$

Because the proposed PSD estimators are not of closed form, in general it is not possible to compute their MSE analytically. Instead, we measure the MSE achieved by the considered PSD estimators in an experiment involving a test signal simulating reverberant and noisy speech. Because the proposed estimators lack closed form, we were only able to numerically verify their unbiasedness. Unbiasedness of the estimators from [15] can be shown analytically (proof omitted). As a result, the MSE of all considered PSD estimators is equal to their variance.

In the special case when the input signal contains no additive noise component ($\mathbf{x}(n) = \mathbf{0}$), the proposed PSD estimators and their MSE can be found analytically. For this restricted scenario it is also possible to analytically find the MSE of the estimators in [15]. In Appendix B we show that in the noise-free scenario the MSE of the proposed estimators is always lower than (or equal to) that of the estimators in [15].

4.1 Experimental setup

In the present experiment, the goal was to measure and compare the performance of the considered estimators in a synthetic scenario where all the assumptions made in Section 2 are precisely met. Thus, in each iteration of the experiment a test signal consisting of 25000 STFT sample vectors $\mathbf{y}(n)$, independently drawn from a circularly-symmetric, multivariate complex Gaussian distribution, was used. The covariance matrix of that distribution was modeled according to (C.7) (i.e. simulating a cross-PSD matrix of a reverberant and noisy speech signal) with known and constant ϕ_s and ϕ_r. Component $\mathbf{s}(n)$ was modeled using a realistic RTF vector $\mathbf{d}$, measured in an anechoic chamber using microphones of two hearing aids placed on the ears of a head and torso acoustic simulator (HATS) and a loudspeaker positioned in front of the HATS. Each of the two behind-the-ear hearing aids had two microphones spaced 1 cm

apart, resulting in the total number of microphones $M = 4$. Component $\mathbf{r}(n)$ was modeled using a normalized cross-PSD matrix $\mathbf{\Gamma_r}$ measured in a simulated cylindrically isotropic sound field using the same microphone array as before. The cross-PSD matrix of the component $\mathbf{x}(n)$ was modeled as a scaled identity matrix. Both evaluated algorithms were set to estimate the input covariance matrix (C.8) using the $L = 10$ most recent time frames.

The simulation experiment was repeated for two different conditions. In the first one, the MSE of the PSD estimation was evaluated as a function of frequency, and the values of ϕ_s and ϕ_r were fixed to result in a speech-to-reverberation ratio (SRR) of 0 dB (averaged over all microphones). In the second condition, the MSE was evaluated as a function of the SRR and the frequency was fixed to 1500 Hz. In both conditions, the additive noise component $\mathbf{x}(n)$ was scaled such that its power was 10 dB lower than the power of the component $\mathbf{r}(n)$ (averaged over all microphones).

4.2 Experimental results

Results obtained in the two described conditions are presented in Figures C.2a and C.2b, respectively. These results are complemented by Cramér-Rao lower bounds (CRLBs) which set a theoretical bound on the lowest possible variance any unbiased estimator of $\phi_s(n)$ and $\phi_r(n)$ can achieve in the considered signal model (C.7). We outline the derivation of the CRLBs in Appendix C.

From Figures C.2a and C.2b it may be observed that in all experimental conditions the target speech and the late reverberation PSD estimators proposed in this study (labeled as "Proposed") achieve lower MSE than the corresponding estimators from [15] ("Braun"). The difference between the MSEs yielded by the late reverberation PSD estimators was substantial. However, the difference between the two target speech PSD estimators was very small in virtually all conditions. This was expected because the two target speech estimators are conditioned on different late reverberation PSD estimators but are otherwise identical [17].

As shown in Figure C.2a, the late reverberation PSD estimator by Braun and Habets achieved MSEs close to the CRLB only for frequencies below 1 kHz. For higher frequencies the MSE of estimation was up to 3.5 dB higher than the CRLB. The proposed late reverberation PSD estimator achieved MSEs close to the CRLB at all analyzed frequencies and SRRs. It is worth highlighting that this has been accomplished despite the simplifications (C.14)–(C.19) of the signal model and the likelihood function used in the derivation of the proposed estimator. It follows, that even the exact MLEs defined using the unmodified signal model (C.11), or any other unbiased estimator based on (C.11), could at best perform only slightly better that the proposed simplified method. The steep rise of the MSEs and the CRLBs for low frequencies is due to the wavelength becoming much larger than the dimensions of the array. This results in an increasing correlation of the microphone signals, which limits the attainable gain from averaging between the microphones.

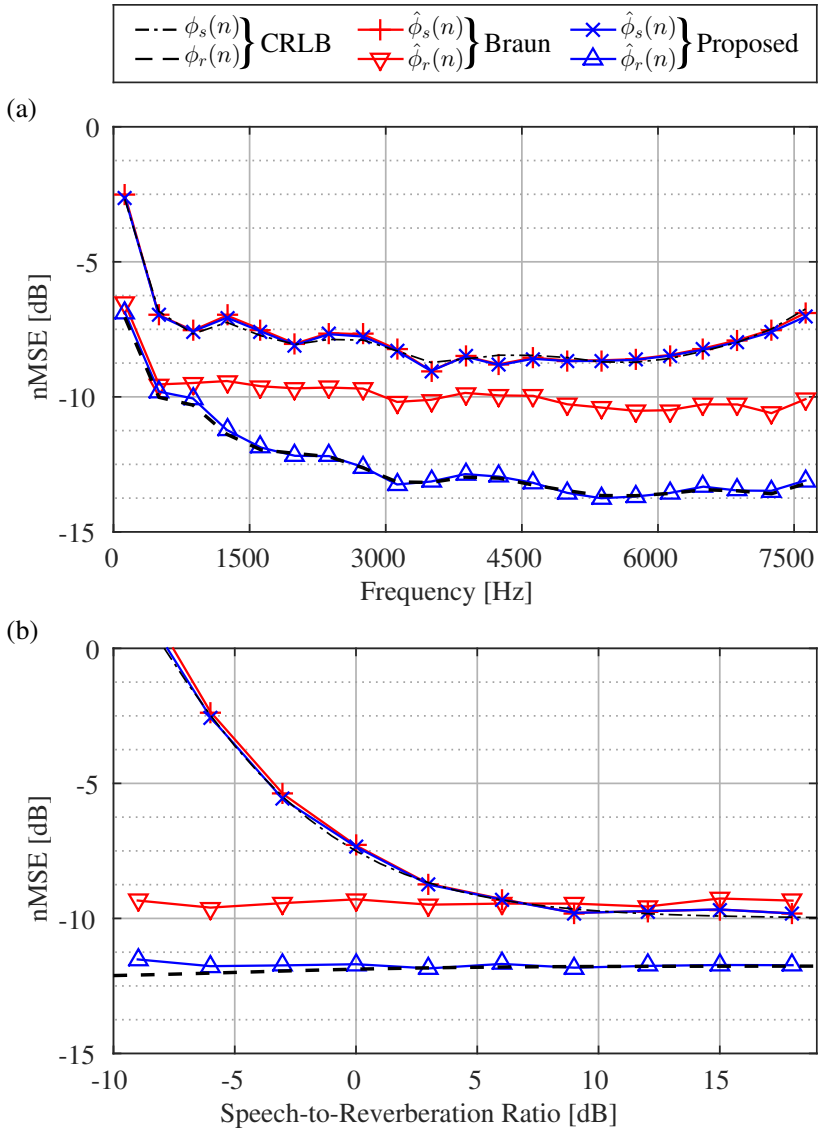


Fig. C.2: Normalized MSE of PSD estimation of the proposed PSD estimators (Proposed) and the PSD estimators from [15] (Braun) as a function of: (a) frequency (SRR: 0 dB), (b) SRR (frequency: 1500 Hz). $M = 4$, $L = 10$.

The performance difference between the two compared late reverberation PSD estimators is substantial despite the fact that both estimators are derived using the maximum likelihood method and are based on similar signal models. The specific cause of this difference is the likelihood function used in [15]. This likelihood is based on the assumption that real and imaginary parts of all entries of the blocked sample cross-PSD matrix $\mathbf{B}^H\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)\mathbf{B}$ are mutually independent Gaussians with equal variances. However, since sample covariance matrices are Hermitian, entries that are symmetric with respect to the main diagonal are complex conjugate pairs (and, hence, not independent). Furthermore, the distribution of diagonal elements of sample covariance matrices has a positive support (i.e. they are not Gaussian), and, generally, the elements of sample covariance matrices can have different variances. In the proposed method the likelihood function (C.19) is defined directly on the (modified) input signal STFT vector and a more realistic assumption on its PDF. Despite the simplifications of the signal model, this results in nearly optimal performance.

As expected, and as can be observed in Figure C.2b, negative SRRs resulted in a much higher target speech PSD estimation MSE (and CRLB) than positive SRR values. Because both "Braun" and "Proposed" late reverberation PSD estimators are based on the blocked version of the input signal, their theoretical performance does not depend on the target speech component and, hence, the SRR.

5 Evaluation of an MWF based on the proposed PSD estimator: objective performance measures

The proposed PSD estimator and the estimator in [15] are both primarily intended for use with an MWF for joint speech dereverberation and denoising. Therefore, it is of interest to evaluate the influence the PSD estimators have on the performance of the MWF. To this end, we conducted an experiment where realistically simulated reverberant and noisy speech signals were processed by the MWF based on either the proposed or the competing PSD estimator from [15]. The speech dereverberation and denoising performance of the two versions of the MWF was measured and compared in terms of the frequency-weighted segmental SNR (FWSegSNR) [18], perceptual evaluation of speech quality (PESQ) [19] measure, mean noise attenuation (NA), mean reverberation attenuation (RA), and speech-to-speech-distortion ratio (SNR-S) [20, 21].

5.1 Experimental setup

Both versions of the MWF were implemented as a concatenation of an MVDR beamformer and a single-channel Wiener post-filter. The MVDR beamformer

Table C.1: Basic acoustic parameters of the reverberant conditions

Room	T_{60} [s]	C_{50} [dB]	DRR [dB]
Bathroom	0.8	5.2	−10.1
Cellar	1.2	5.7	2.2
Staircase	2.3	11.0	4.1
Office	1.4	8.8	2.3
Auditorium	1.3	13.4	5.2
Isotropic	1.0	4.7	−0.4

coefficients $\mathbf{w}_{\mathrm{MVDR}}(n)$ were calculated according to (C.13) with the estimate of the total interference cross-PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{v}}(n)$ based on $\hat{\phi}_r(n)$. The output signal $\hat{s}(n)$ of the MWF was computed as:

$$\hat{s}(n) = \left[\frac{\hat{\phi}_{s_o}(n)}{\hat{\phi}_{s_o}(n) + \hat{\phi}_{v_o}(n)}\right] \mathbf{w}_{\mathrm{MVDR}}^H(n)\mathbf{y}(n),$$

where:

$$\hat{\phi}_{s_o}(n) = \hat{\phi}_s(n),$$
$$\hat{\phi}_{v_o}(n) = \mathbf{w}_{\mathrm{MVDR}}^H(n)\hat{\mathbf{\Phi}}_{\mathbf{v}}(n)\mathbf{w}_{\mathrm{MVDR}}(n),$$

denote the estimated PSDs of the target speech and the total interference at the output of the MVDR beamformer, respectively.

Contrary to the experiment in Section 4, in this experiment the goal was to compare the performance of the estimators in realistic conditions (violating some of the assumptions made in Section 2) and for a practical application (in hearing aids). Thus, the microphone signals were generated using real speech recordings from the TIMIT database [40] and several reverberant and noisy conditions based on real room impulse responses (RIRs) and simulated microphone noise. Specifically, we used a subset of the TIMIT database containing 17 minutes of male and female speech. TIMIT sentences were convolved with RIRs measured in five real rooms using a microphone array composed of two behind-the-ear hearing aids on the HATS (same as described in Section 4). The reverberation time T_{60}, clarity index C_{50}, and the direct-to-reverberation ratio (DRR) of these five RIRs are presented in Table C.1. The rooms are denoted by their function as: "Bathroom", "Cellar", "Staircase", "Office", and "Auditorium", and represent a wide range of acoustic conditions a hearing aid user might encounter. A sixth, synthetic impulse response, where the reverberation was modeled as perfectly cylindrically isotropic was also used and is denoted as "Isotropic". To simulate the electrical noise that is generated by real-world microphones, spatially white and spectrally pink noise was added to the convolved speech signals. The simulations were repeated for two levels of

that noise, such that at the frequency of 1 kHz the noise PSD was either 20 dB or 30 dB lower than the PSD of the target speech material.

The sampling frequency of the simulated time-domain signals was 16 kHz and the STFT length was set to 8 ms ($T = 128$ samples). This ensured a processing delay of the MWF shorter than 10 ms, which is a requirement for hearing aid systems. A square root Hann window with 50% overlap between frames was used in the analysis filterbank and in the overlap-add inverse STFT procedure used for re-synthesis of the output signal. The input cross-PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)$ was estimated using recursive averaging (equivalent to exponential weighting) with a time constant of 50 ms (instead of the moving average smoothing used in (C.8)). For processing of the signals simulated using each of the six impulse responses, the MWF algorithm and the PSD estimators were implemented using RTF vectors $\mathbf{d}$ extracted from the first 2.5 ms of the RIR in question (i.e. $t_s = 40$ samples). For the RIRs used in the experiment this resulted in $\mathbf{d}$ being based solely on the direct path response. It follows that the early reflections (particularly strong in the "Bathroom" condition) were left unaccounted for in the assumed signal model. This resulted in a realistic mismatch between the used RTF vector $\mathbf{d}$ and the actual RTF of the target speech component in the simulated signals. Moreover, because $\mathbf{d}$ depended only on the direction of the target source, the assumption that $\mathbf{d}$ is known became more realistic. The normalized cross-PSD matrix $\mathbf{\Gamma}_{\mathbf{r}}$ of the cylindrically isotropic sound field was measured *a priori* in a simulated cylindrically isotropic sound field using the same microphone array as used for measuring the RIRs. In none of the five real rooms, the late reverberation was truly isotropic which, again, resulted in a realistic mismatch between the assumed model and the actual structure of the signal. Only in the "Isotropic" condition the model of the target signal and of the reverberation component was accurate.

5.2 Experimental results

The results of the experiment are presented in Figure C.3. Performance scores obtained by using the MWF based on the proposed PSD estimator ("Proposed") and on the estimator proposed in [15] ("Braun") are included along the scores obtained by using only the MVDR part of the two MWFs ("MVDR"). The scores calculated from the unprocessed input signal ("Input") are included for reference. The relative performance between the proposed and the competing MWFs and MVDRs was the same for the higher and the lower microphone noise level setting. Thus, we show only the results obtained for the -30 dB setting, which better corresponds to the typical microphone noise and speech levels encountered in practice.

In all simulated conditions, both versions of the MWF and the MVDR beamformer succeeded in improving FWSegSNR and PESQ. The RA was also always positive, indicating algorithms' effectiveness in reducing the reverberation. However, the NA scores were exclusively negative, indicating that *on average* all algorithms amplified the noise. This was expected because the NA

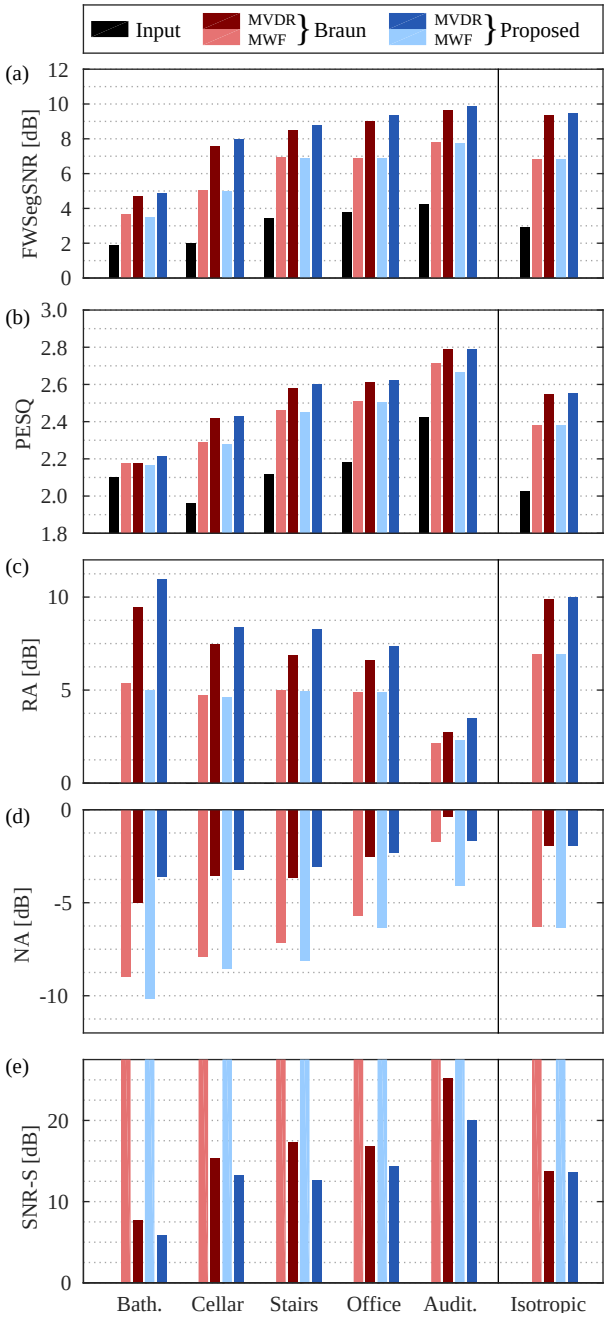


Fig. C.3: (a) FWSegSNR, (b) PESQ, (c) RA, (d) NA, and (e) SNR-S scores obtained by using MWFs and MVDR beamformers based on the PSD estimators from [15] (denoted "Braun") and the proposed estimators (denoted "Proposed"). Scores obtained from the unprocessed input signal ("Input") are also included.

measure (as well as RA and SNR-S) only accounts for those signal segments where the target speech component is active (see [21]). Because the MVDR beamformer adapts to jointly suppress the noise and the late reverberation, it was expected that during speech and, hence, reverberation activity the noise component will have negligible impact on the MVDR coefficients. Naturally, during speech and reverberation absence the MVDR beamformers adapted to primarily reduce the noise component.

The total improvement of the FWSegSNR and PESQ over the unprocessed signal was greatest in the "Isotropic" and lowest in the "Bathroom" condition. This difference can be explained by the fact that in the "Isotropic" condition $\mathbf{\Gamma_r}$ and $\mathbf{d}$ accurately characterized the actual input signal whereas in the "Bathroom" condition the input signal did not match the assumed model. Prominent early reflections present in the "Bathroom" condition were unaccounted for and resulted in substantial leakage of the early speech component into the output of the blocking matrix. This lead to an overestimation of the late reverberation PSD and, ultimately, over-suppression and distortion of the target speech in the post-filter (note the very high RA and very low SNR-S in this condition).

The differences in the performance scores obtained by using the MVDR beamformers based on "Braun" and "Proposed" PSD estimators were very small. Although the performance difference between the MWFs was somewhat bigger, it was still only moderately large. For example in the "Isotropic" condition "Proposed" MWF performed only marginally better than "Braun". In the remaining conditions the difference is larger and the proposed method appears to systematically perform better than "Braun". This suggests that the proposed estimators are more robust to the mismatch between the signal model and its actual structure than the estimators from [15]. The SNR-S measure indicated stronger speech distortion when using the proposed PSD estimator. While being a clear disadvantage, low SNR-S scores are counterbalanced by higher RA and NA values.

Informal listening tests indicated similar trends as the objective performance measures. In all simulated conditions, the MWFs resulted in a decrease of the perceived reverberation and noise strength. The MVDR beamformers also reduced the amount of perceived interference, but to a smaller degree. Differences between "Braun" and "Proposed" MWFs were barely perceivable; only in specific signal scenarios a small increase in the audibility of musical noise could be noticed in the "Braun" MWF output. This was expected because the PSD estimators from [15] have higher MSE.

We close this section by noting that (when implemented in Matlab) the proposed algorithm resulted in computation times roughly 1.7 times longer than the algorithm from [15].

6 Evaluation of an MWF based on the proposed PSD estimator: speech intelligibility improvement

In addition to the two experiments with technical/objective performance measures in Sections 4 and 5, we conducted a speech intelligibility (SI) test with human subjects. Dantale II [41] sentences were presented via Sennheiser HD280 pro headphones to 20 subjects, who were requested to select the words they heard from an on-screen list of options [42].

6.1 Experimental setup

Stimuli were constructed as follows. The Dantale II sentences were concatenated with 2 s of silence before and after the utterance and underwent the same realistic reverberation simulation as in the "Cellar" condition in Section 5, corresponding to a frontal position of the target source at a distance of 2 m. Since the SI in this condition was close to 100%, speech-like interference consisting of randomly shifted and superimposed copies of the international speech test signal (ISTS) [43] was added to the reverberated Dantale II sentences. The interferer signals were convolved with 5 RIRs recorded in the same room as the target RIR but with the sound source positioned at 90°, 135°, 180°, −135°, and at −90° azimuth angle, at 2 m distance. Each of the simulated babble talkers radiated the same power as the target source. Different levels of SI were achieved by manipulating the DRR of the the target source RIR. This was done by attenuating the direct part of the target speech while keeping the rest of the signal intact. In this way the DRR was offset by 0, −4, −8, and −12 dB from its original value of 2.2 dB (cf. Table C.1).

The RTF vector $\mathbf{d}$ and the cross-PSD matrix $\mathbf{\Gamma_r}$ were obtained in the same way, and the simulated microphone signals were processed using the same algorithms as in Section 5. The additional noise cross-PSD matrix $\mathbf{\Phi_x}$ was estimated from the first 2 s of each stimuli, which was known to contain only the reverberated ISTS babble and the simulated microphone noise. In order to provide correct binaural cues of the target speech, signals presented to each of the subjects' ears were processed by separate instances of the algorithms, each using the front microphone of the corresponding hearing aid as the reference position. In the unprocessed condition ("Input") the signals of the left and right reference microphones were presented to the corresponding ears of the subject. This allowed the subjects to localize the target and the ISTS interferers at their original (simulated) positions and benefit from the binaural advantage [44]. In the processed conditions this was not possible, as all components of the enhanced signals were perceived as coming from the target direction (a known side-effect of using binaural beamformers [45]). To each of the experimental conditions five Dantale II sentences were randomly assigned (independently for each subject). The sentences were processed and then presented to subjects in

a randomized order.

6.2 Experimental results

The word intelligibility obtained in each of the processing conditions was calculated as the percentage of words identified correctly by the subjects and is plotted in Fig. C.4 as a function of the DRR offset. In order to interpret these results, we performed a two-way repeated measures ANOVA procedure [46] on the rationalized arcsine-corrected [47] subject mean word intelligibility scores. The effect of the processing type ($F_{4,76} = 232.6$), the DRR offset ($F_{3,57} = 383.8$), and the interaction term ($F_{12,228} = 5.0$) on the measured intelligibility were all found to be significant at the $p < 0.001$ level. Pairwise comparisons of the marginal means revealed that: *a)* each of the algorithms significantly improved the SI over the "Input", *b)* the MWFs outperformed their corresponding MVDR beamformers, and additionally, *c)* the "Proposed MWF" outperformed the "Braun MVDR". The familywise type I error rate was limited to 1% using Bonferroni correction.

The lack of significant differences between the SI obtained with the proposed and the competing PSD estimators was somewhat expected, given the minute instrumental performance differences of the two MWFs and MVDR beamformers obtained in Section 5. On the other hand, significant improvement of SI resulting from the post-filters of the two MWFs is apparently in contrast with the general understanding that single channel spectral filters usually fail to increase SI [48]. The fact that the post-filters of the two evaluated MWFs succeeded in improving SI can be explained by the fact that they were computed using information from the multi-microphone signal (contrary to the single channel schemes discussed in [48]).

7 Conclusion

In this paper we have proposed a pair of novel ML-based speech and late reverberation PSD estimators. The proposed method models the interference as consisting of late reverberation and additive noise; in this sense it can be seen as an extension of the method in [16] which only considers the late reverberation. We have numerically demonstrated that the proposed estimator yields lower mean squared error (MSE) of PSD estimation than the method in [15], and that this MSE is very close to the corresponding Cramér-Rao lower bound.

In an experiment with realistically simulated reverberation, we have compared speech dereverberation performance of an MWF based on the proposed estimator and on the estimator from [15]. The proposed estimator generally resulted in higher FWSegSNR, PESQ, RA, and NA scores than the estimator from [15]. However, the SNR-S indicated stronger speech distortion. In terms of speech intelligibility, the MWFs based on both PSD estimators provided statistically significant improvements over the unprocessed signal, but were not

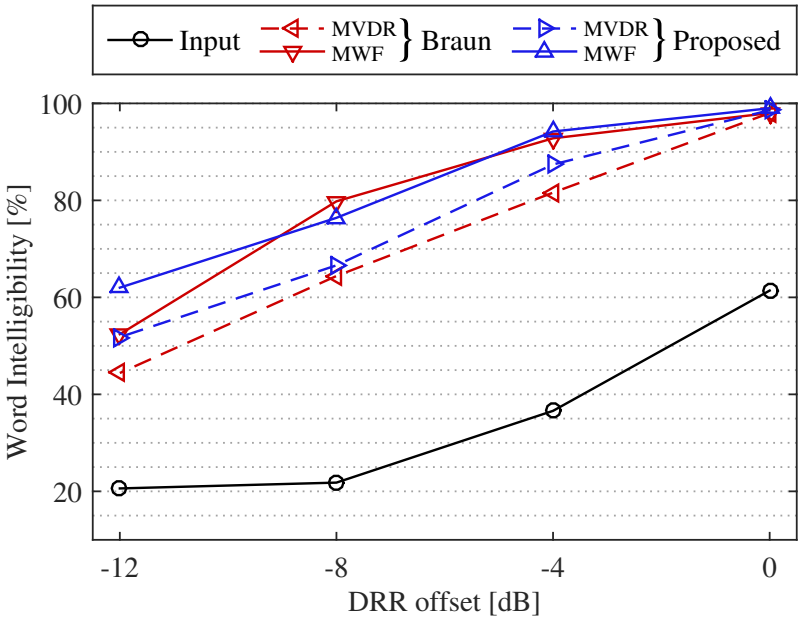


Fig. C.4: Word intelligibility scores obtained in the listening test with the RIR from the "Cellar" condition and ISTS interferers, averaged across 20 subjects.

significantly different from each other. The output of both MWFs was statistically significantly more intelligible than the output of the corresponding MVDR beamformers.

Evaluation of the proposed algorithm in environments which more severely violate the assumptions made in this paper is an area for future work. In an ongoing study, we evaluate the proposed algorithm's robustness to erroneous estimates of the direction of the target speech arrival. Preliminary results for signals without the noise component have already been published in [49].

8 Acknowledgements

The authors would like to thank the reviewers for their thorough and insightful review of this manuscript, and Asger Heidemann Andersen, Oticon A/S, for allowing the use of his software implementation of the Dantale II matrix test.

A Properties of the proposed late reverberation PSD estimator

In the following, we show that in the majority of practical cases the polynomial equation (C.22) (repeated below for convenience) has exactly one real-valued

root and that this root is non-negative. Due to this property, the proposed MLE of $\phi_r(n)$ can be found more easily, as the likelihood (C.19) does not need to be calculated in order to determine which of the roots of (C.22) corresponds to the MLE of $\phi_r(n)$.

$$p(\phi_r) = \sum_{m=1}^{M-1} p_m(\phi_r), \quad \text{where} \tag{C.22}$$

$$p_m(\phi_r) = \underbrace{\left(\phi_r - \frac{g_m(n)-1}{\lambda_{\mathbf{\Gamma},m}}\right)}_{\text{order } 1} \underbrace{\prod_{k=1}^{M-1,\, k\neq m} \left(\phi_r + \frac{1}{\lambda_{\mathbf{\Gamma},k}}\right)^2}_{\text{order } 2(M-2)}.$$

We begin by noting that the order of the polynomial $p(\phi_r)$ depends linearly on the number of microphones M and it is equal to $2M-3$. Because this is always an odd number, at least one root of $p(\phi_r)$ is real. This means that for all possible input signals the proposed method will return a result.

The polynomial $p(\phi_r)$ is a sum of $M-1$ polynomials $p_m(\phi_r)$, and each $p_m(\phi_r)$ has exactly one root of algebraic multiplicity one and exactly $M-2$ roots of multiplicity two (cf. (C.22)). The double roots of each $p_m(\phi_r)$ are equal to $-\lambda_{\mathbf{\Gamma},k}^{-1}$. These roots are always negative because $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$ is assumed positive-definite, i.e. all of its eigenvalues $\lambda_{\mathbf{\Gamma},m}$ are strictly positive. The singular root of each $p_m(\phi_r)$ is equal to $(g_m(n)-1)\lambda_{\mathbf{\Gamma},m}^{-1}$, which is non-negative if and only if $g_m(n) \geq 1$. This condition is expected to be satisfied whenever $\phi_r(n) \geq 0$, because (C.17), (C.18):

$$\begin{aligned} E[g_m(n)] &= E\big\{[\mathbf{U}^H \hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\mathbf{U}]_{m,m}\big\} \\ &= \phi_r(n)\lambda_{\mathbf{\Gamma},m} + 1. \end{aligned} \tag{C.26}$$

The structure of the component polynomials $p_m(\phi_r)$ allows us to draft their approximate plots in Figure C.5. We note that each of the component polynomials attains a value of zero, but it does not cross it at the double roots. The $M-1$ double roots are repeated between $p_m(\phi_r)$ but each of them is absent from exactly one of the polynomials (cf. (C.22)). It follows that the polynomial $p(\phi_r)$ is strictly negative between $-\infty$ and the lowest of the singular roots. Analysis of the derivatives and inflection points of $p(\phi_r)$ leads to a conclusion that given $g_m(n) \geq 1$, the graph of the polynomial $p(\phi_r)$ crosses the abscissa only once at a point between the lowest and the highest of the singular roots of the component polynomials, i.e. it has exactly one real root and it is non-negative. The condition $g_m(n) \geq 1$ can be expected to be satisfied, because in reverberant scenarios $\phi_r(n)$ is almost always positive (cf. (C.26)). Our simulations confirm that; the polynomial (C.22) has a single positive root in over 99% of cases.

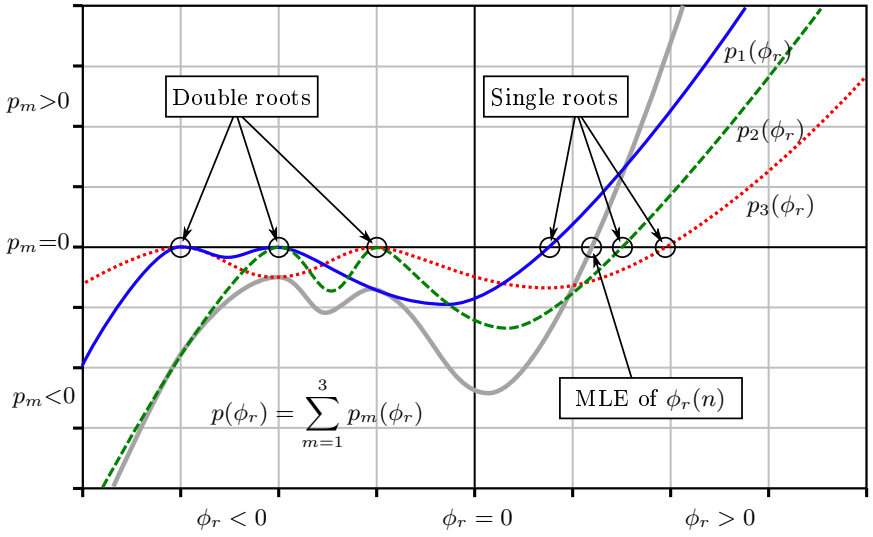


Fig. C.5: Schematic illustration of the polynomial (C.22) (denoted $p(\phi_r)$) and its $M-1$ components $p_m(\phi_r)$ for $M=4$.

B Theoretical performance of the proposed late reverberation PSD estimator in noise absence

In this appendix we compare analytical expressions for the mean squared error (MSE) of the PSD estimators proposed in this study and the PSD estimators proposed by Braun and Habets in [15]. This comparison does not appear to be possible in the general case where $\mathbf{x}(n) \neq \mathbf{0}$ because of the lack of a closed-form solution for the proposed late reverberation PSD estimator. Therefore, in this appendix we are restricted to the special case where no additive noise component is present (i.e. $\mathbf{x}(n) = \mathbf{0}$). As shown in Section 3.3, in such a signal scenario the proposed late reverberation PSD estimator can be written in closed-form (C.23).

Since the proposed PSD estimators in the special case of $\mathbf{x}(n) = \mathbf{0}$ are identical to the speech and reverberation PSD estimators proposed by us in [16], the comparison we make in this appendix is equivalent to the one presented in [17]. We outline it in the following for completeness.

The target speech PSD estimator proposed by Braun and Habets in [15] has the same form as the target speech PSD estimator (C.12) proposed in the present study. The difference between the estimators is that they are conditioned on different late reverberation PSD estimates. Hence, it is sufficient to compare the late reverberation PSD estimators in order to capture the difference between the two PSD estimation methods.

We start the comparison of the late reverberation PSD estimator proposed by Braun and Habets (denoted $\hat{\phi}_{r,\text{Braun}}(n)$) and the one proposed in this study (denoted $\hat{\phi}_{r,\text{Kukl.}}(n)$) by noting that they are both unbiased (proof omitted):

$$E\big[\hat{\phi}_{r,\text{Kukl.}}(n)\big] = \phi_r(n), \quad E\big[\hat{\phi}_{r,\text{Braun}}(n)\big] = \phi_r(n).$$

Therefore, the MSEs of these estimators are identical to their variances.

The variance of the proposed late reverberation PSD estimator (C.23) can be shown to be equal to (for proof see [14]):

$$\text{var}\big(\hat{\phi}_{r,\text{Kukl.}}(n)\big) = \phi_r^2(n)\frac{1}{L}\frac{1}{M-1}. \tag{C.27}$$

The variance of the late reverberation PSD estimator proposed by Braun and Habets [15] has been previously derived in [17] and can be concisely written as:

$$\text{var}\big(\hat{\phi}_{r,\text{Braun}}(n)\big) = \phi_r^2(n)\frac{1}{L}\frac{1}{M-1}\left(1+\frac{\tilde{\gamma}^2}{\bar{\gamma}^2}\right), \tag{C.28}$$

where $\tilde{\gamma}$ and $\bar{\gamma}$ denote the sample variance and the mean of the squared eigenvalues of the matrix $\mathbf{\Gamma}_{\check{\mathbf{r}}} = \mathbf{B}^H\mathbf{\Gamma}_{\mathbf{r}}\mathbf{B}$, respectively.

Comparing (C.28) and (C.27) and using the fact that $\tilde{\gamma}$ and $\bar{\gamma}$ are non-negative we can conclude that the MSE of $\hat{\phi}_{r,\text{Braun}}(n)$ can be either greater or equal to the MSE of $\hat{\phi}_{r,\text{Kukl.}}(n)$, but can never be lower. The MSEs of these two estimators are equal only when the eigenvalues of $\mathbf{\Gamma}_{\check{\mathbf{r}}}$ are all equal (i.e. when $\tilde{\gamma}^2 = 0$). Since $\mathbf{\Gamma}_{\check{\mathbf{r}}}$ is Hermitian, it follows that for this special case to occur, $\mathbf{\Gamma}_{\check{\mathbf{r}}}$ must be a scaled identity matrix [50]. In all other cases, the proposed late reverberation PSD estimator has lower MSE than the one from [15].

An important observation is that for $M = 2$ the matrix $\mathbf{\Gamma}_{\check{\mathbf{r}}}$ reduces to a scalar, such that $\tilde{\gamma}^2$ is always equal to zero. It follows that for $M = 2$ the proposed late reverberation PSD estimator (C.23) and the one proposed by Braun and Habets [15] achieve the same MSE. In this case they are, in fact, identical (proof omitted).

C Cramér-Rao lower bounds on PSD estimation

In this appendix we outline the calculation of the Cramér-Rao lower bounds (CRLBs) included in Figures C.2a and C.2b. By definition, the CRLBs are equal to the elements of the inverse of the Fisher information matrix (FIM). The i,j-th element of the FIM is defined as follows [51]:

$$\mathcal{I}_{i,j} = -E\left[\frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\theta_i\,\partial\theta_j}\right], \tag{C.29}$$

where $\mathcal{L}$ is the log-likelihood function of the parameter vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T$, given the input data. For a p-parameter signal model the FIM is a $p \times p$ symmetric matrix. For L independent identically distributed circularly-symmetric

complex Gaussian observations the i,j-th element of the FIM is found as [51]:

$$\mathcal{I}_{i,j} = L \operatorname{tr}\left[\boldsymbol{\Phi}_{\mathbf{y}}^{-1} \frac{\partial \boldsymbol{\Phi}_{\mathbf{y}}}{\partial \theta_i} \boldsymbol{\Phi}_{\mathbf{y}}^{-1} \frac{\partial \boldsymbol{\Phi}_{\mathbf{y}}}{\partial \theta_j}\right], \tag{C.30}$$

where $\boldsymbol{\Phi}_{\mathbf{y}}$ is the cross-PSD matrix of the input signal. Note that because of the above equation, any invertible linear operation applied to the input signal vector (such as whitening) does not change the FIM or the CRLB.

In the signal model considered in this study (C.7) there are two unknown parameters ($\boldsymbol{\theta} = [\phi_s, \phi_r]^T$); hence, the FIM is a 2×2 matrix. Using the log-likelihood function (C.10) in (C.29), or equivalently the cross-PSD matrix (C.7) in (C.30) we obtain:

$$\mathcal{I} = \begin{bmatrix} \mathcal{I}_{ss} & \mathcal{I}_{rs} \\ \mathcal{I}_{sr} & \mathcal{I}_{rr} \end{bmatrix} \tag{C.31}$$

$$\mathcal{I}_{ss} = L \operatorname{tr}\left[\boldsymbol{\Phi}_{\mathbf{y}}^{-1} \mathbf{d}\mathbf{d}^H \boldsymbol{\Phi}_{\mathbf{y}}^{-1} \mathbf{d}\mathbf{d}^H\right], \tag{C.32}$$

$$\mathcal{I}_{rr} = L \operatorname{tr}\left[\boldsymbol{\Phi}_{\mathbf{y}}^{-1} \boldsymbol{\Gamma}_{\mathbf{r}} \boldsymbol{\Phi}_{\mathbf{y}}^{-1} \boldsymbol{\Gamma}_{\mathbf{r}}\right], \tag{C.33}$$

$$\mathcal{I}_{rs} = \mathcal{I}_{sr} = L \operatorname{tr}\left[\boldsymbol{\Phi}_{\mathbf{y}}^{-1} \boldsymbol{\Gamma}_{\mathbf{r}} \boldsymbol{\Phi}_{\mathbf{y}}^{-1} \mathbf{d}\mathbf{d}^H\right]. \tag{C.34}$$

Similarly to the proposed PSD estimators, the CRLBs do not appear to be possible to be derived analytically in the general case. For the special case when $\mathbf{x}(n) = \mathbf{0}$, closed-form expressions for the CRLBs can be derived (see e.g. [14]). When $\mathbf{x}(n) \neq \mathbf{0}$ the FIM can be inverted numerically and (by definition) the CRLBs are obtained as:

$$CRLB(\phi_s) = [\mathcal{I}^{-1}]_{1,1}, \tag{C.35}$$

$$CRLB(\phi_r) = [\mathcal{I}^{-1}]_{2,2}. \tag{C.36}$$

The CRLBs included in Figures C.2a and C.2b were calculated using (C.30)–(C.36) and normalized by the squared parameter of interest (analogous to the normalization of MSEs in (C.25)).

References

[1] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am.*, vol. 113, no. 6, pp. 3233–3244, 2003.

[2] P. A. Naylor, "Introduction," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. Springer, 2010.

[3] D. Schmid *et al.*, "Variational bayesian inference for multichannel dereverberation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 8, pp. 1320–1335, Aug 2014.

[4] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multichannel equalization," in *14th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sept 2014, pp. 139–143.

[5] J. Benesty, S. Makino, and J. Chen, "Introduction," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Berlin, Germany: Springer, 2005, ch. 1, pp. 1–8.

[6] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec 1979.

[7] P. C. Loizou, *Speech Enhancement: Theory and Practice.* Taylor & Francis, 2007.

[8] K. Lebart, J. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[9] E. A. P. Habets, "Single-channel speech dereverberation based on spectral subtraction," in *15th Annu. Workshop Circuits, Systems, Signal Process.*, 2004, pp. 250–254.

[10] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912–915, 1977.

[11] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 5, Apr 1988, pp. 2578–2581.

[12] S. Doclo *et al.*, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. R. Liu, Eds. Wiley, 2008, pp. 269–302.

[13] ——, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7-8, pp. 636–656, Jul.–Aug. 2007.

[14] J. Jensen and M. S. Pedersen, "Analysis of beamformer-directed single-channel noise reduction system for hearing aid applications," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brisbane, Australia, 2015, pp. 5728–5732.

[15] S. Braun and E. A. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. 21st Eur. Signal Process. Conf. (EUSIPCO)*, Marrakech, Morocco, 2013, pp. 1–5.

[16] A. Kuklasiński *et al.*, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Proc. 22nd Eur. Signal Process. Conf. (EUSIPCO)*, Lisbon, Portugal, 2014, pp. 61–65, **(Paper A in this thesis)**.

[17] ——, "Multi-channel PSD estimators for speech dereverberation – a theoretical and experimental comparison," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brisbane, Australia, 2015, pp. 91–95, **(Paper B in this thesis)**.

[18] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan 2008.

[19] "Perceptual evaluation of speech quality: an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Rec. P. 862*, 2001.

[20] S. Gustafsson *et al.*, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 245–256, Jul 2002.

[21] J. S. Erkelens *et al.*, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.

[22] I. Cohen, "Speech enhancement using a noncausal a priori snr estimator," *IEEE Signal Process. Lett.*, vol. 11, no. 9, pp. 725–728, 2004.

[23] G. W. Elko, E. Diethorn, and T. Gänsler, "Room impulse response variation due to temperature fluctuations and its impact on acoustic echo cancellation," in *Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Kyoto, Japan, 2003, pp. 67–70.

[24] J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *J. Sound and Vibration*, vol. 102, no. 2, pp. 217–228, 1985.

[25] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 204–207, JUL 2003.

[26] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, 2005.

[27] J. Jensen, I. Batina, R. C. Hendriks, and R. Heusdens, "A study of the distribution of time-domain speech samples and discrete fourier coefficients," in *Proc. SPS-DARTS*, vol. 1, 2005, pp. 155–158.

[28] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug 2001.

[29] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, 2004.

[30] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 7, pp. 1420–1434, 2009.

[31] H. Kuttruff, *Room Acoustics*, 5th ed. Taylor & Francis, 2009.

[32] R. K. Cook *et al.*, "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 27, no. 6, pp. 1072–1077, 1955.

[33] G. W. Elko, "Spatial coherence functions for differential microphones in isotropic noise fields," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001, pp. 61–85.

[34] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 5, pp. 1072–1077, 2010.

[35] H. Ye and R. D. DeGroat, "Maximum likelihood DOA estimation and asymptotic Cramér-Rao bounds for additive unknown colored noise," *IEEE Trans. Signal Process.*, vol. 43, no. 4, pp. 938–949, 1995.

[36] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 35, no. 10, pp. 1365–1376, 1987.

[37] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Bucharest, Romania, 2012, pp. 295–299.

[38] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Nov. 2012.

[39] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, 2007.

[40] J. S. Garofolo *et al.*, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*. NIST, 1993.

[41] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *Int. J. Audiology*, vol. 42, no. 1, pp. 10–17, 2003.

[42] E. R. Pedersen and P. M. Juhl, "Speech in noise test based on a ten-alternative forced choice procedure," *Joint Baltic-Nordic Acoust. Meeting*, 2012.

[43] I. Holube *et al.*, "Development and analysis of an international speech test signal (ISTS)," *Int. J. Audiology*, vol. 49, no. 12, pp. 891–903, 2010.

[44] B. C. J. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[45] S. Doclo *et al.*, "Extension of the multi-channel wiener filter with itd cues for noise reduction in binaural hearing aids," in *IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, 2005, pp. 70–73.

[46] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press, 2011.

[47] G. A. Studebaker, "A rationalized arcsine transform," *J. Speech, Language, Hearing Res.*, vol. 28, no. 3, pp. 455–462, 1985.

[48] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am.*, vol. 122, no. 3, pp. 1777–1786, 2007.

[49] A. Kuklasiński *et al.*, "Multi-channel Wiener filter for speech dereverberation in hearing aids – sensitivity to DoA errors," in *60th Audio Eng. Soc. Int. Conf.*, Leuven, Belgium, 2016, **(Paper D in this thesis)**.

[50] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[51] S. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, ser. Prentice Hall Signal Processing Series. Prentice-Hall PTR, 1993.

Paper D

Multi-channel Wiener filter for speech dereverberation in hearing aids – sensitivity to DoA errors

A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen

The paper has been presented at the
60th Audio Engineering Society International Conference: Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS),
Leuven, Belgium, 2016.

Abstract

In this paper we study the robustness of a recently proposed Multi-channel Wiener Filter-based speech dereverberation algorithm to errors in the assumed direction of arrival (DoA) of the target speech. Different subsets of microphones of a pair of behind-the-ear hearing aids are used to construct various monaural and binaural configurations of the algorithm. Via a simulation experiment with frontally positioned target it is shown, that when correct DoA is assumed binaural configurations of the algorithm almost double the improvement of PESQ measure over monaural configurations. However, in conditions where the assumed DoA is increasingly incorrect, the performance of the binaural configurations is shown to deteriorate more quickly than that of the monaural configurations. In effect, for large DoA errors it is the simpler, monaural configurations that perform better.

* * *

Paper E

Contralateral microphones in multi-channel Wiener filters for hearing aids – benefits and tradeoffs

A. Kuklasiński and J. Jensen

The paper has been submitted to the
Journal of Audio Engineering Society

Abstract

We consider an adaptive multi-channel Wiener filter (MWF) for joint dereverberation and noise reduction in hearing aids. Using STOI and FWSegSNR measures, we compare bilateral and binaural configurations of this MWF for: (a) *different directions of arrival (DoAs) of the target speech,* (b) *different errors in the assumed target DoA, and* (c) *different levels of the microphone self-noise. The binaural MWFs, while being much less robust against DoA errors, are found to outperform the bilateral MWF if the correct DoA is assumed. Furthermore, the bilateral MWF is shown to be affected by the microphone self-noise more than the binaural MWFs. Finally, the advantage of the binaural over the bilateral MWF is demonstrated through a speech intelligibility test with reverberant and noisy speech stimuli.*

* * *

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-762-1

AALBORG UNIVERSITY PRESS